Registration-Free Learnable Multi-View Capture of Faces in Dense Semantic Correspondence
Supplementary Material

Panagiotis P. Filntisis[1,3,4] George Retsinas[1] Radek Daneček[4]
Vanessa Sklyarova[4, 5] Petros Maragos[1,2,3] Timo Bolkart[6]

[1]Institute of Robotics, Athena Research Center, 15125 Maroussi, Greece
[2]School of ECE, NTUA, Greece [3]HERON – Hellenic Robotics Center of Excellence, Athens, Greece
[4]MPI for Intelligent Systems, Tübingen, Germany [5] ETH Zurich, Zurich, Switzerland [6] Google, Switzerland

This supplementary material provides additional details and results for MOCHI.

1. Implementation Details

FaMoS Setup The FaMoS dataset acquisition setup comprises both grayscale stereo pairs (8 pairs) and RGB cameras (8 cameras). For training MOCHI, we rely exclusively on the 8 RGB cameras; an example of the input views from this configuration is shown in Fig. 1. To ensure a fair comparison, we also retrained the baseline method TEMPEH[1] using its official open-source implementation, configuring it to use the same set of 8 RGB cameras instead of the stereo pairs. For more information on FaMoS please see [1].

Figure 1. **Multi-view input setup.** Example views from the 8 RGB cameras used in the FaMoS dataset. We utilize only these RGB streams for both MOCHI and the TEMPEH baseline.

Exclusion of Scalp Region from Evaluation In the quantitative evaluation presented in the main text, we explicitly exclude the scalp region from our evaluation protocol. This exclusion is necessary because subjects in the FaMoS dataset wore hair caps (or hair nets) to conceal their hair during capture. These caps introduce noise into the reconstructed raw scans—often capturing the geometry of the fabric rather than the subject's actual anatomical shape. Consequently, the raw scans in this region do not serve as a reliable ground truth for modeling the scalp. We illustrate examples of this in Fig. 2.

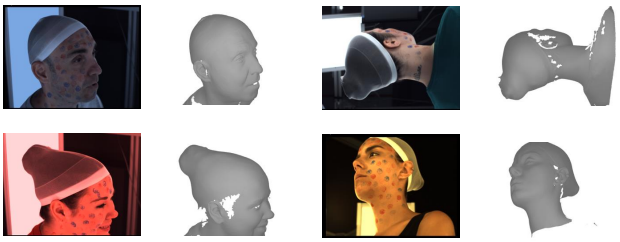

Figure 2. **Impact of hair caps on ground-truth scans.** We show example input images and their corresponding raw scans from the FaMoS dataset. Depending on the fit of the hair cap, the resulting scans exhibit significant distortion in the scalp region, often including the geometry of the cap itself. As a result, this region does not faithfully represent the subject's actual scalp and is excluded from our quantitative metrics.

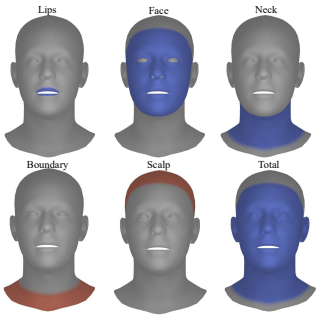

Figure 3. **Visualization of evaluation masks.** We utilize standard FLAME vertex masks to compute regional metrics. **Blue regions** indicate vertices included in the evaluation: (Top Row) Lips, Face, and Neck; (Bottom Right) the combined Head mask. **Red regions** indicate excluded vertices: (Bottom Left) the lower boundary and (Bottom Center) the scalp region, which are removed to ensure robust evaluation against capture noise and to ignore regions non-relevant to face reconstruction.

Evaluation Regions. We adopt the official vertex masks provided with the FLAME model [6] to define our regions of interest. Fig. 3 visualizes the specific vertex subsets used

Name	Full w/o Scalp			Face			Lips			Neck		
	Median ↓	Mean ↓	Std ↓	Median ↓	Mean ↓	Std ↓	Median ↓	Mean ↓	Std ↓	Median ↓	Mean ↓	Std ↓
w/ p2s t=5	0.16	0.31	1.61	0.14	0.25	1.52	0.23	0.47	1.23	0.27	0.64	1.31
w/ p2s t=10	0.13	0.27	1.58	0.11	0.21	1.50	0.18	0.38	1.15	0.22	0.56	1.24
w/ p2s t=20	0.10	0.23	1.57	0.09	0.18	1.50	0.13	0.28	0.99	0.18	0.49	1.16
w/ p2s t=50	0.07	0.18	1.55	0.07	0.15	1.49	0.08	0.19	0.89	0.15	0.42	1.08
w/ p2s t=100	**0.06**	**0.16**	1.54	0.06	0.14	1.49	0.06	**0.15**	**0.85**	0.13	0.39	1.04
w/o normals t=5	0.14	0.31	1.60	0.12	0.23	1.47	0.21	0.47	1.27	0.23	0.58	1.29
w/o normals t=10	0.11	0.26	1.57	0.10	0.20	1.44	0.15	0.39	1.22	0.19	0.49	1.16
w/o normals t=20	0.09	0.22	1.55	0.08	0.17	1.43	0.10	0.31	1.19	0.16	0.43	1.08
w/o normals t=50	0.07	0.19	1.52	0.06	0.15	1.41	0.07	0.23	1.13	0.13	0.39	1.04
w/o normals t=100	**0.06**	0.17	1.51	**0.05**	**0.13**	1.40	0.06	0.20	1.10	**0.12**	**0.37**	**1.03**
no-pretrain t=5	1.13	1.46	1.49	1.06	1.31	1.14	1.21	1.58	1.59	1.66	2.61	2.97
no-pretrain t=10	1.02	1.38	1.50	0.94	1.23	1.15	1.12	1.55	1.63	1.51	2.54	3.03
no-pretrain t=20	0.75	1.16	1.47	0.68	1.02	1.12	0.89	1.36	1.58	1.12	2.23	2.94
no-pretrain t=50	0.44	0.94	1.46	0.39	0.81	**1.10**	0.56	1.12	1.56	0.69	1.85	2.80
no-pretrain t=100	0.32	0.89	1.48	0.28	0.77	1.15	0.38	1.02	1.56	0.5	1.67	2.73
TTO direct opt t=5	0.17	0.36	1.57	0.16	0.27	1.41	0.25	0.60	1.37	0.30	0.97	1.91
TTO direct opt t=10	0.16	0.33	1.56	0.15	0.26	1.41	0.24	0.56	1.34	0.28	0.92	1.90
TTO direct opt t=20	0.12	0.29	1.55	0.10	0.23	1.41	0.19	0.50	1.31	0.23	0.85	1.88
TTO direct opt t=50	0.09	0.26	1.53	0.08	0.22	1.41	0.10	0.40	1.23	0.18	0.76	1.88
TTO direct opt t=100	0.08	0.28	1.54	0.07	0.26	1.46	0.09	0.37	1.14	0.16	0.82	2.17
MOCHI TTO (full model) t=5	0.14	0.30	1.60	0.12	0.23	1.49	0.20	0.46	1.27	0.23	0.57	1.26
MOCHI TTO (full model) t=10	0.11	0.26	1.58	0.09	0.20	1.46	0.14	0.37	1.22	0.19	0.48	1.14
MOCHI TTO (full model) t=20	0.08	0.22	1.56	0.08	0.17	1.45	0.10	0.30	1.19	0.15	0.42	1.08
MOCHI TTO (full model) t=50	0.07	0.18	1.53	0.06	0.14	1.43	0.07	0.23	1.13	0.13	0.38	1.04
MOCHI TTO (full model) t=100	**0.06**	**0.16**	1.52	**0.05**	**0.13**	1.42	**0.05**	0.19	1.09	**0.12**	**0.37**	**1.03**
Classic Registrations	0.10	0.24	**1.44**	0.08	0.16	1.21	0.09	0.36	1.39	0.24	0.90	2.39

Table 1. **Quantitative ablation of Test-Time Optimization (TTO).** We evaluate the impact of different loss functions, supervision signals, and optimization strategies on the FaMoS validation set. We report the **Median**, **Mean**, and **Std** of the point-to-surface error (mm) for the full head (excluding scalp) and specific facial regions. *Direct Opt* refers to initializing with MOCHI and optimizing vertices directly—rather than fine-tuning the parameters of the refinement model—while *No-Pretrain* optimizes the network from random initialization. Our full model (MOCHI TTO) outperforms all baselines and surpasses the quality of classical registrations.

in our quantitative analysis. The regions highlighted in **blue** (Lips, Face, Neck, and the combined Head) are used for calculating reconstruction errors. Conversely, the regions highlighted in **red** are excluded from evaluation. Specifically, we omit the scalp region due to the capture artifacts discussed previously, and the lower mesh boundary that are not relevant to facial reconstruction quality and can be affected by non-relevant to the head scan noise.

2. Extended Ablation Studies

Extra Qualitative Comparisons. Figure 4 presents additional qualitative results on FaMoS validation and test subjects, comparing MOCHI (with and without TTO at 50 iterations) against TEMPEH [1] and classic registrations. We show a broad range of examples and expressions, further demonstrating MOCHI's accuracy and generalization.

Test-Time Optimization (TTO) Strategies. In Table 1, we provide the complete set of numerical results corresponding to Fig. 7 in the main text. Furthermore, we introduce the **No-Pretrain** baseline where the local refinement module is randomly initialized and optimized from scratch for each scan, instead of being initialized from pre-trained MOCHI weights.

Without MOCHI initialization (No-Pretrain), performance degrades significantly, underlining the importance of learned priors.

Additionally, Fig. 6 shows how reconstruction quality improves progressively from step 5 to 100.

Direct Vertex Optimization Fig. 5 illustrates the robustness of our TTO scheme. Under identical conditions, direct vertex optimization leads to high-frequency artifacts, while our latent TTO produces smooth, topologically valid surfaces.

Chamfer Distance While prior methods [1, 5] primarily rely on point-to-surface error, the Chamfer distance is another prevalent metric in 3D vision. However, similar to point-to-surface error, the Chamfer distance relies on non-differentiable nearest-neighbor selection, which can cause artifacts and can make training more unstable.

In Fig. 7, we compare the training dynamics of the coarse stage over 500 iterations using three different objectives: the conventional point-to-surface loss, the Chamfer loss,

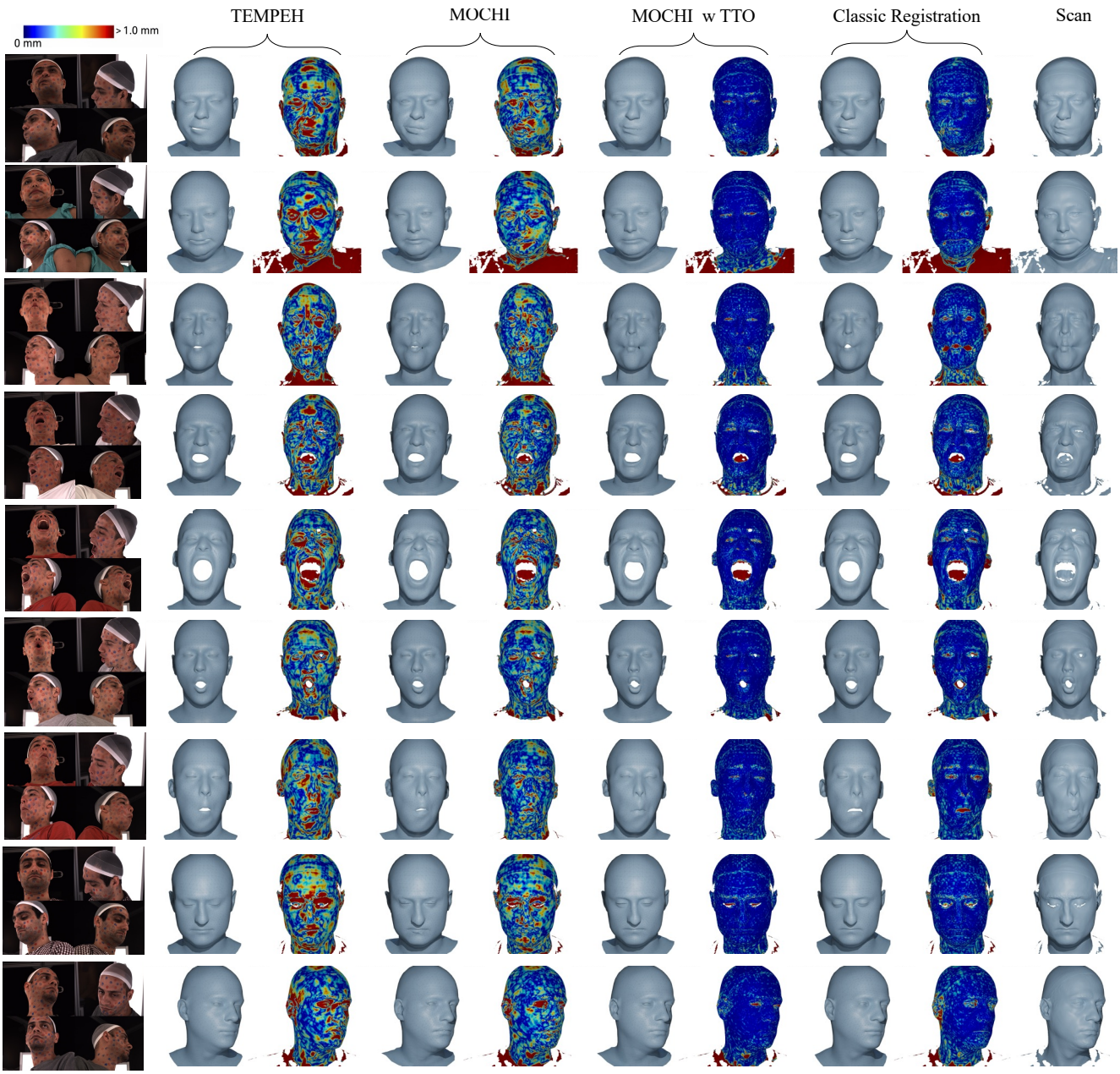

Figure 4. **Qualitative results on FaMoS validation and test sets (we show 4 of 8 input views).** From Left to Right: *TEMPEH* [1], *MOCHI*, *MOCHI + TTO* (50 iters), *Classic Registration*, and the *Scan*. For each method we show the predicted mesh (left) and the point-to-surface error heatmap on the scan (right; red $\geq$ 1.0 mm, lower is better). MOCHI improves over TEMPEH on unseen subjects even without using registrations for training; adding TTO further reduces error and can surpass the classic manual registration pipeline.

and our differentiable point-map loss. As the weight of the point-to-surface loss increases, reconstructions rapidly develop visible artifacts and self-intersections, effectively bypassing the topology enforcement of our regularizers. Similarly, while the Chamfer distance avoids the severe topological collapses seen with point-to-surface loss, it still induces artifacts and noise (e.g., in the ears and eyeballs). In contrast, supervision with point-maps exhibits significantly smoother behavior, preserving clean geometry and stable gradients even at higher loss weights.

Limitations We visualize typical failure cases of our method in Fig. 8. While MOCHI TTO generally refines geometry, it can encounter difficulties when the raw scan contains teeth, which are not represented in the FLAME

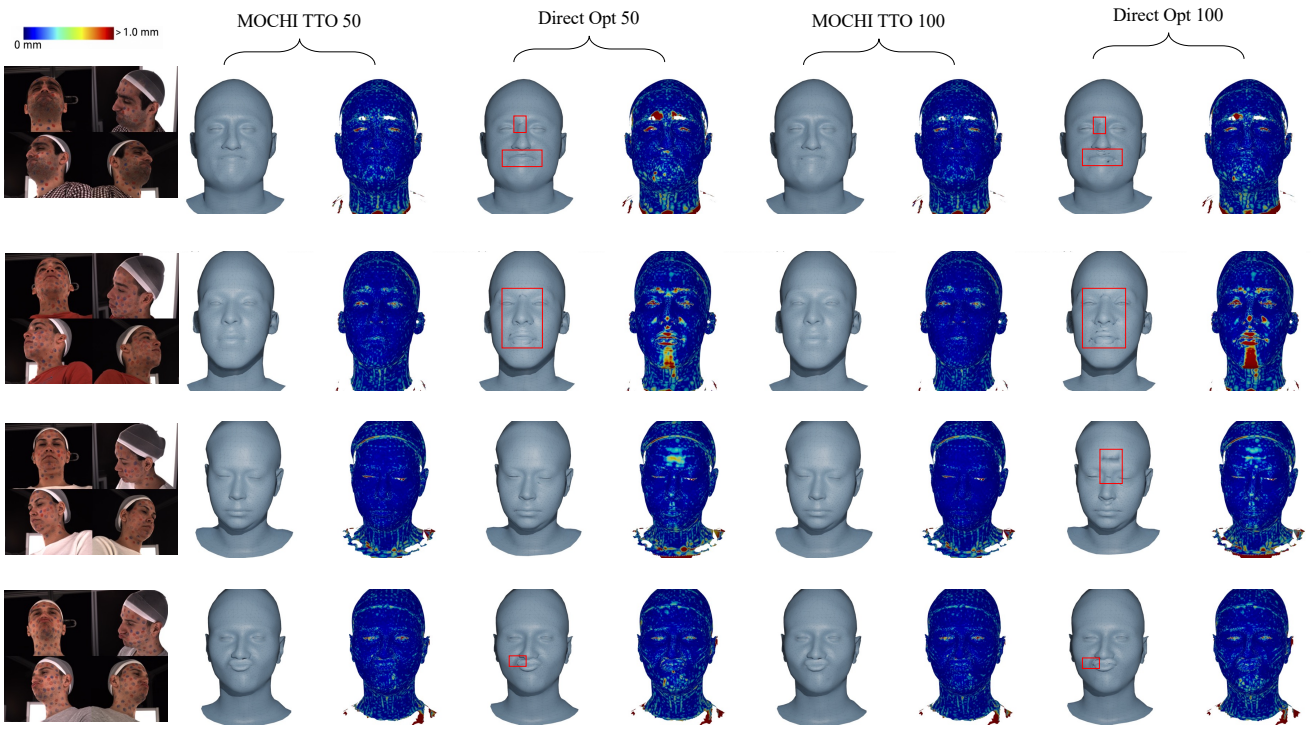

Figure 5. **Robustness of optimization strategies.** We compare our latent-space optimization (MOCHI TTO) against directly optimizing mesh vertices (Direct Opt) at 50 and 100 iterations. While Direct Opt reduces the numerical point-to-surface error, it is prone to artifacts, spikes, and surface irregularities (highlighted by red boxes). In contrast, MOCHI TTO leverages the network's learned prior, ensuring the reconstruction remains smooth and anatomically plausible while accurately fitting the target scan.

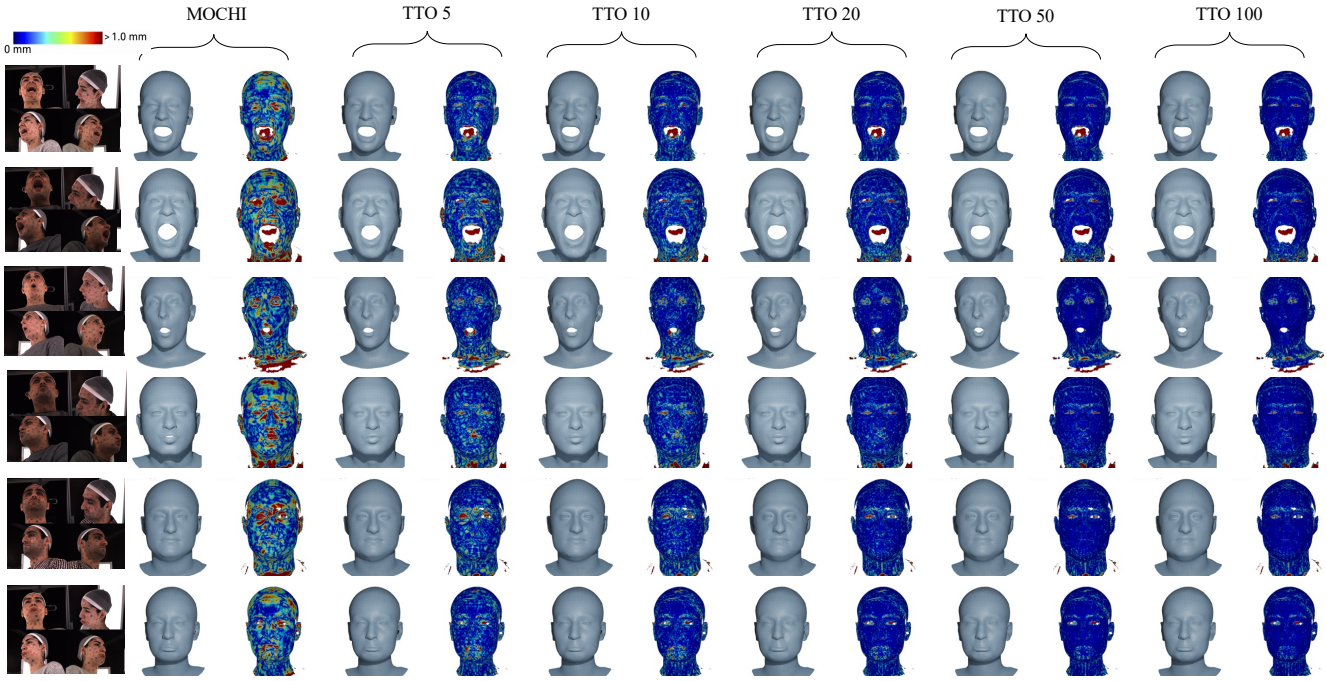

Figure 6. **Qualitative progression of Test-Time Optimization.** This figure shows the evolution of mesh refinement over 100 iterations on example images from FaMoS test set.

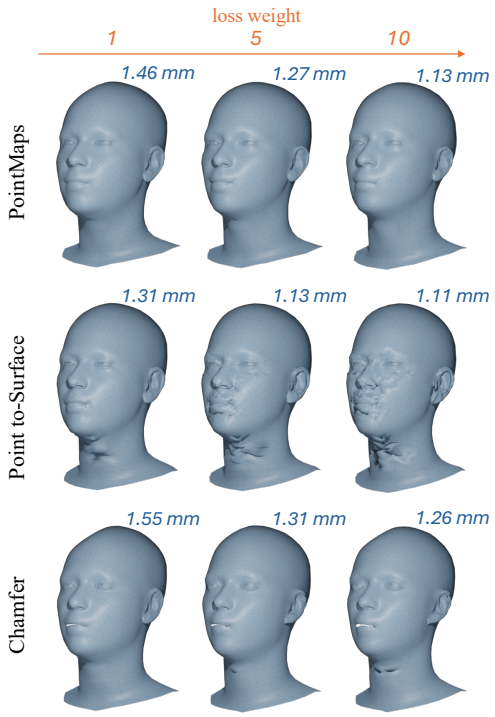

Figure 7. **Impact of loss type and weight.** We visualize the training of the coarse model using point maps (top), point-to-surface loss (middle), and Chamfer loss (bottom). When starting from a coarse prediction initialized only with landmark pretraining, increasing the point-to-surface loss weight leads to severe artifacts. While Chamfer distance results in fewer artifacts, it can still introduce errors (e.g., in the ears); in contrast, point-map supervision remains stable and topology-consistent. Blue numbers indicate the point-to-surface error (mm) relative to the ground-truth scan.

model. Additionally, large head poses can occasionally lead to artifacts in the base model. Such cases can happen more often when we train MOCHI without 2D supervision from the predicted landmarks of the dense landmark tracker.

3. Additional Dataset Evaluation

FaceScape. To further validate generalization, we evaluate MOCHI on FaceScape [10], which uses a different capture rig with per-sample self-calibration via structure-from-motion. This results in inconsistent camera parameters and arbitrary scale across samples. To evaluate on FaceScape, we align scans to a canonical frame and fine-tune our model. Since the official registrations are also unaligned, we rigidly align them to the scans for fair comparison. Qualitative results comparing MOCHI, MOCHI-TTO, and the official FaceScape registration are shown in Fig. 9. As can be seen, MOCHI-TTO produces registrations that are more accurate than the official FaceScape registrations, confirming that our method generalizes beyond the FaMoS capture setup.

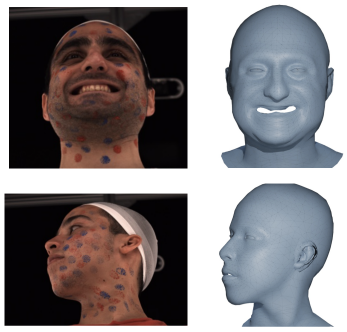

Figure 8. **Failure cases. Top Row (MOCHI TTO):** Since the FLAME topology lacks explicit teeth geometry, test-time optimization may incorrectly distort the lips by pulling them inward to minimize the point-to-surface distance to the teeth present in the raw scan. **Bottom Row (MOCHI):** In cases of large head poses, the base model may occasionally present with artifacts.

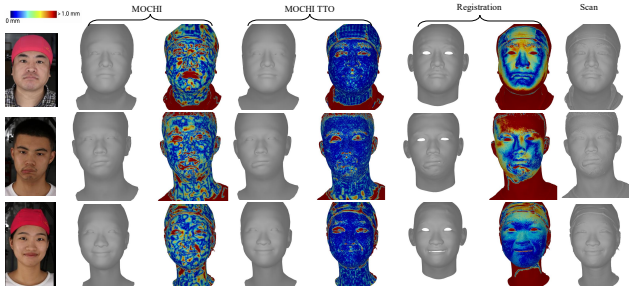

Figure 9. **Qualitative results on FaceScape.** From left to right: *MOCHI*, *MOCHI TTO*, the official *FaceScape registration*, and the *Scan*. For each method we show the predicted mesh (left) and the point-to-surface error heatmap (right; red $\geq$ 1.0 mm, lower is better).

4. Dense Landmark Tracker

4.1. Synthetic Data With Blender

Motivation. As described in the Method section of the main paper, our approach relies on an accurate dense landmark tracker for supervision. There are several ways to obtain high-quality dense landmark datasets:

(1) fitting 3D morphable models (3DDMs) to in-the-wild images, which is limited by the fact that unconstrained face reconstruction is still an open problem;

(2) deriving landmarks from registrations of high-quality 3D face scans captured in controlled studio environments, which we explicitly aim to avoid, as our goal is not to depend on registrations;

(3) synthetically generating data, which offers full control over the facial geometry, appearance, and deformation, and therefore provides perfect ground truth.

Given these considerations, we adopt the third option.

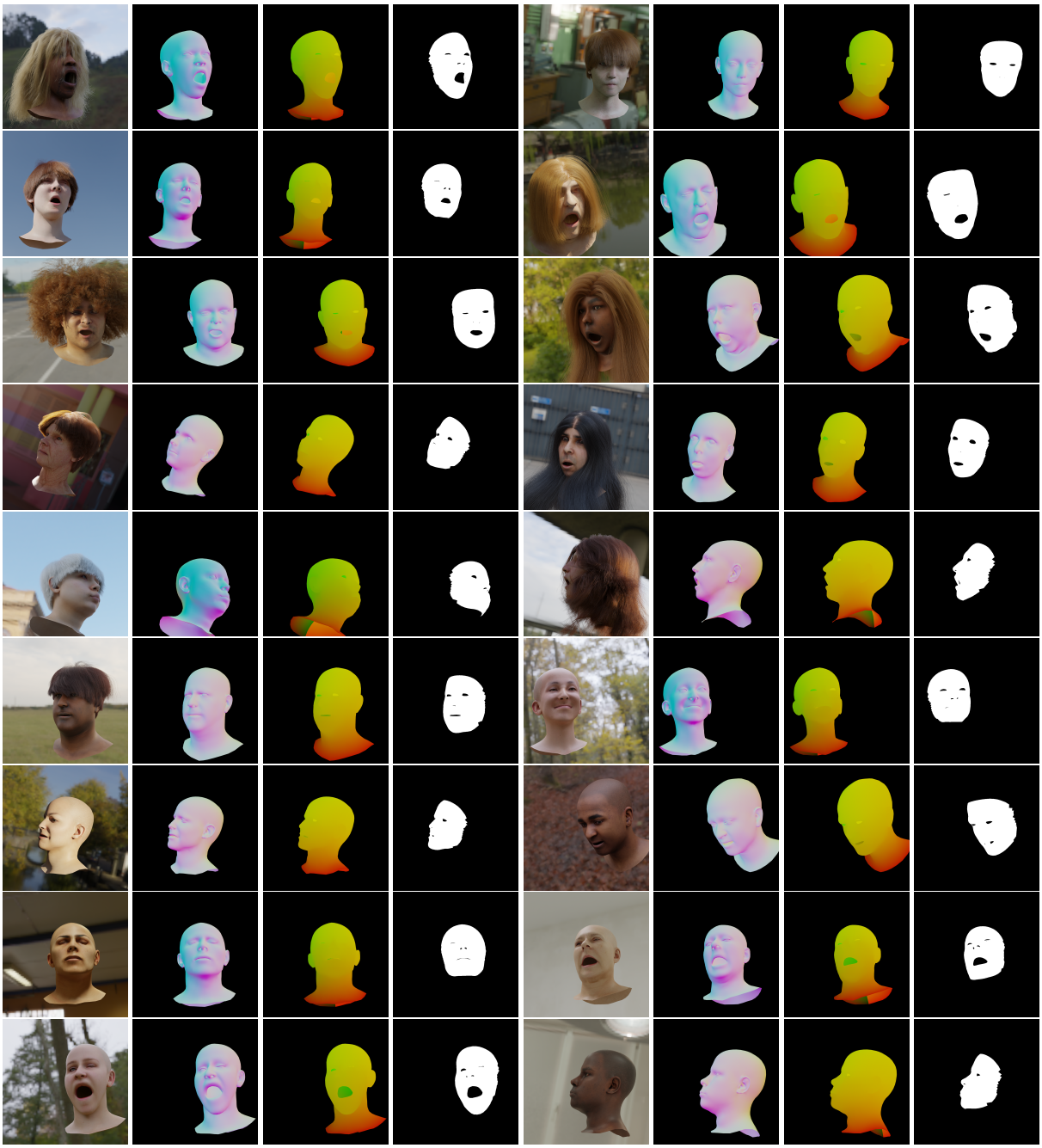

Figure 10. **Synthetic data.** From left to right: Path-traced RGB image, normal map, uv-map, binary face mask. We encourage the reader to zoom in on this figure in its digital form.

Generating Geometry. To generate the geometry we randomly sample a FLAME identity, a FLAME expression, FLAME jaw opening angle, eyeball rotation and the activation of a separate eye-blinking blendshapes. To generate hair, we use HAAR [9] to sample 150 various hairstyles. For each generated face, we either keep it bald, or sample one of the hairstyles.

Albedo. We sample 54 (26 female and 28 male) very high quality textures purchased from 3D Scan Store, which cover all different skin tones and eye colors. These textures were re-topologized to be compatible with the FLAME UV-map. For each rendering, we uniformly sample one texture.

Illumination. For illumination, we employ high quality HDRI environment maps obtained from PolyHaven. They were captured in a variety of indoor and outdoor scenes and cover many possible illumination settings. For each rendering, we first sample one of 665 environment maps. Then, we randomly rotate the environment map about the vertical axis.

Camera position. We randomly sample the camera position to be in the frontal hemisphere of the rendered face. We also slightly vary the "look-at" point such that is not always looking at the center of the face. Furthermore, we perturb the camera distance and the focal length.

Output. We use Blender [2] to render all the images. First, the Cycles path tracer renders the RGB image of the scene generated with the process described above. Then we remove hair (if the sample had any) and switch to the EEVEE renderer to render UV-map, normal map, and flame regions masks. The resulting images along with the mesh and camera parameters, provide perfect GT to train our dense landmark tracker. Figure 10 shows a few example outputs.

Limitations. While our data is sufficient for our tasks, it does have a few limitations. First, FLAME models the face and neck only, which results in a floating-head appearance. Although this is not realistic, we do not observe any negative impact on model performance. Second, we do not model the mouth cavity, teeth, or tongue, as these components are not included in FLAME. Finally, our synthetic data does not include facial hair or accessories such as glasses, which may lead to reduced performance when these are present. Nevertheless, for our primary use case, namely landmark tracking for multi-view facial capture, subjects typically do not exhibit such variations, so this limitation has minimal practical impact.

4.2. 2D Landmark Tracker

Using the previously described synthetic dataset, we train a dedicated 2D landmark tracker. We employ a Vision Transformer (ViT-Large) backbone [3], initialized with DINOv3 weights [8]. To maintain the robust semantic features learned during pre-training while adapting to our task, we freeze the backbone parameters and inject Low-Rank Adaptation (LoRA) [4] modules into the attention and feed-forward blocks. Specifically, we apply LoRA with rank $r = 8$ and $\alpha = 16$ to the query, key, value, and projection matrices in the attention layers, as well as the fully connected layers in the MLP blocks. The features are aggregated via global average pooling and passed to a lightweight regression head consisting of a LayerNorm and a final Linear projection that predicts the 2D normalized coordinates of the FLAME mesh vertices.

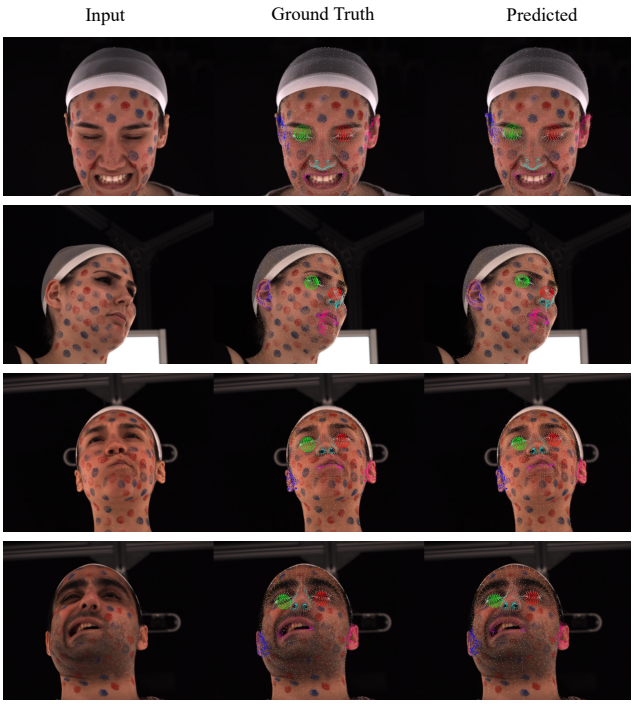

Figure 11. **Qualitative evaluation of the 2D landmark tracker on the FaMoS dataset.** From left to right: the input RGB image, the Ground Truth landmarks (obtained by projecting the available registration), and the landmarks predicted by our tracker. The vertices are color-coded by semantic region (e.g., green/red for eyes, pink for lips, blue for ears). Despite being trained exclusively on synthetic data, the tracker generalizes robustly to real-world images, accurately capturing dense correspondence across challenging expressions and poses.

Training and Evaluation. We train the tracker for 100 epochs on our generated synthetic dataset. The input images are resized to 512×512 pixels. We use the AdamW optimizer [7] with a learning rate of 10^{-4} and weight decay of 10^{-4}. The learning rate is scheduled using a cosine annealing strategy with a linear warm-up. The model is supervised using a Mean Squared Error (MSE) loss between the predicted 2D coordinates and the ground-truth projected vertices from the synthetic rendering. To improve generalization to real-world data, we apply aggressive geometric and photometric image augmentations during training.

To validate our 2D tracker, we perform a qualitative comparison against projected FLAME registrations (from a classical pipeline) on a subset of the FaMoS dataset. Figure 11 visualizes the output of our tracker—trained exclusively on synthetic data—when applied to real-world samples. In our MOCHI pipeline, these 2D predictions serve

as semantic regularizers, anchoring difficult regions such as the eyes and lips. As such, the tracker does not need to achieve pixel-perfect accuracy, but rather provide stable, semantically meaningful guidance.

References

[1] Timo Bolkart, Tianye Li, and Michael J. Black. Instant multi-view head capture through learnable registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 2, 3

[2] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. 7

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. 7

[4] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. 7

[5] Jing Li, Di Kang, and Zhenyu He. Grape: Generalizable and robust multi-view facial capture. *arXiv preprint arXiv:2407.10193*, 2024. 2

[6] Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. In *ACM SIGGRAPH Asia 2017*, 2017. 1

[7] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. 7

[8] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025. 7

[9] Vanessa Sklyarova, Egor Zakharov, Otmar Hilliges, Michael J. Black, and Justus Thies. Text-conditioned generative model of 3d strand-based human hairstyles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4703–4712, 2024. 6

[10] Haotian Yang, Hao Zhu, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao. Facescape: A large-scale high quality 3d face dataset and detailed riggable 3d face prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 5