Supplementary Material - SHeaP: Self-supervised Head Geometry Predictor Learned via 2D Gaussians

Anonymous ICCV submission

Paper ID 2158

1. Implementation Details

1.1. Model Hyperparameters and Training Details

3DMM Parameters Estimator The ViT is initialized with FaRL [9], which uses exactly the same network structure as CLIP's [6] VIT-B16 architecture. On an NVIDIA RTX 3090 GPU, this ViT can predict 3DMM parameters at 230 frames per second.

Gaussian Regressor The UV Map Generator uses the Lightweight GAN architecture [4]. We set its latent dimension to 1024 and feature map inverse coefficient parameter to 13. We also remove 3 last blocks of the network such that it outputs a UV region features map $\boldsymbol{M}$ of size 128×128. The graph convolutional network follows a ResNet-like architecture, with 4 residual blocks.

Training Details As described in the main paper, our model is trained on a mix of Nersemble and VFHQ. We entirely exclude the subjects used in the Nersemble geometric evaluation from the training dataset. We train our model for 100,000 steps on a single NVIDIA L40S GPU, with a batch size of 8, which takes around 24 hours to complete.

Loss Weights For the photometric losses, our best model configuration uses:

$$\mathcal{L}_{LI} = 6$$

$$\mathcal{L}_{perc} = 6$$

$$\mathcal{L}_{ID} = 0.3$$

$$\mathcal{L}_{emo} = 0.1.$$

The 3DMM regularizers w_ψ and w_β are both set to 1.0. We also set the weights on $\mathcal{L}_{normals}$ and $\mathcal{L}_{depth}$ to both be 1.0.

Other hyperparamters We set $t_{densify} = 1$, $n_{prune} = 1$, $n_{densify} = 1$ and $t_{history} = 256$. We regularize the Gaussian offsets with $w_{\boldsymbol{x}} = 0.1$; scales with $w_{\boldsymbol{s}} = 0.01$; and opacities with $w_o = 0.001$.

1.2. Details of AffectNet Training

Following a similar methodology to EMOCA [1], to produce our results reported in Table 3, we fit a 4-layer MLP to map predicted FLAME parameters to AffectNet arousal, valence, and emotion category values.

For every comparison method, we first predict FLAME parameters for all images from the AffectNet dataset. These parameters are then used as inputs for the 4-layer MLP. During training, we specifically utilize the FLAME parameters predicted from the training portion of the dataset, while for evaluation, we employ the parameters from the test portion.

The MLP is trained to predict arousal, valence, and emotion categories by optimizing a multi-objective loss function; specifically, we employ a mean squared error loss for valence and arousal predictions, and a cross-entropy loss for the emotion category classification. We train the MLP for 20 epochs and a batch size of 16. We apply dropout to each layer with a rate of 0.3. We use a weight on the emotion classificaiton loss of 0.7, on the valence loss 0.8, and on the arousal loss 0.3. We set the learning rate to 1e-4 and decay this by a factor of 0.6 every epoch. We also use a small amount of weight decay of 1e-6. The model's performance is evaluated using the test portion of the AffectNet dataset, and its predictive accuracy is compared against existing benchmarks to validate the effectiveness of our approach.

2. Stability of Predicted Identity Coefficients Through Time

In theory, the identity-related parameters predicted by our model should be constant across a single video of the same person. To assess this, we obtain the UV region features map $\boldsymbol{M}$ for all images across 1000 different videos taken from VFHQ. We then fit a PCA to the resulting $256 \times 256 \times 1000$ UV feature vectors. In Figure 1, we visualize the first three components of this PCA space for two different subject videos. The PCA features appear almost identical across multiple input images for the same video, while differing substantially from one video to an-

other. This shows that these features are relatively identity-stable and invariant to the pose and expression of the input image.

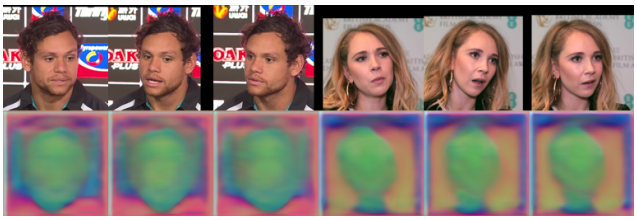

Figure 1. PCA visualization

Figure 2 presents a more quantitative assessment of the stability of identity-related features predicted by our model. We plot the first dimension of the 3DMM shape code β for 100 frames of three different videos, and find that this too is relatively stable through time.

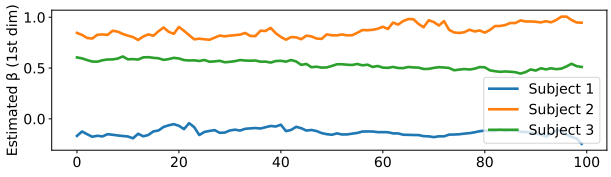

Figure 2. First dimension of β for 3 different subject videos.

3. Qualitative Evaluation on Nersemble dataset

As mentioned in the main text, the NoW dataset [8] is primarily used to assess how effectively methods predict heads with neutral expressions. The true facial mesh is created by using an active stereo system to scan individuals with a neutral expression, guaranteeing precise 3D facial scans. Nevertheless, the mesh obtained this way is not perfectly neutral within the expression space of 3DMM, which is a common target for prediction by most methods. Additionally, rather than solely focusing on the precision of estimated head shapes, our primary concern is the geometric accuracy of the estimated facial expressions, which is vital for numerous downstream applications (e.g., [5]). To achieve this, we developed the Nersemble point cloud dataset, which includes 60 high-fidelity point clouds and 960 images from different perspectives generated from the Nersemble dataset [3].

Since we thoroughly evaluated our method's effectiveness quantitatively on the Nersemble point cloud in the main manuscript, in the following sections, we focus on comparing our method qualitatively with representative single-image head reconstruction methods. We show the error maps by projecting the true point clouds onto their closest points on the head mesh and compute root squared distances for each point. Overall, judging from the error map in Fig. 3, our method outperforms other methods across different capturing angles and with varying expressions.

References

[1] Radek Danecek, Michael J. Black, and Timo Bolkart. EMOCA: Emotion driven monocular face capture and animation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20311–20322, 2022. 1, 3

[2] 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

[3] Tobias Kirschstein, Shenhan Qian, Simon Giebenhain, Tim Walter, and Matthias Nießner. Nersemble: Multi-view radiance field reconstruction of human heads. *ACM Trans. Graph.*, 42(4), 2023. 2

[4] Bingchen Liu, Yizhe Zhu, Kunpeng Song, and Ahmed Elgammal. Towards faster and stabilized gan training for high-fidelity few-shot image synthesis. In *iclr*, 2021. 1

[5] Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20299–20309, 2024. 2

[6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 1

[7] George Retsinas, Panagiotis P. Filntisis, Radek Danecek, Victoria F. Abrevaya, Anastasios Roussos, Timo Bolkart, and Petros Maragos. 3d facial expressions through analysis-by-neural-synthesis. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[8] 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. 2

[9] Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. General facial representation learning in a visual-linguistic manner. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18697–18709, 2022. 1

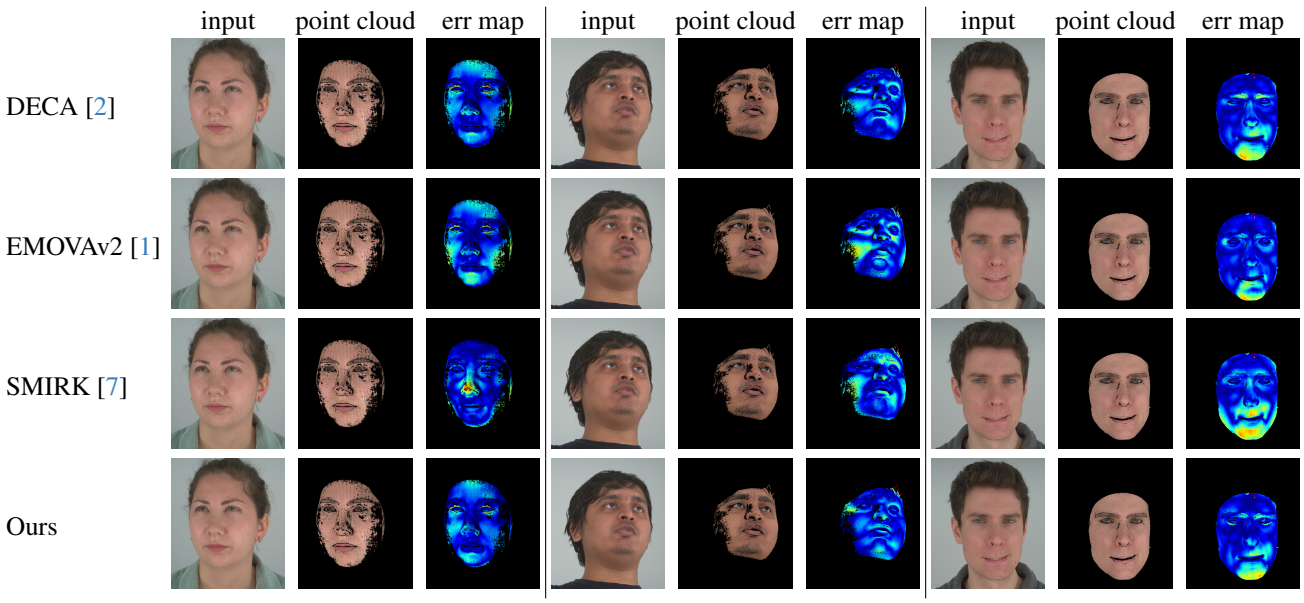

Figure 3. Error maps derived by calculating the root squared Euclidean distance from point clouds to their nearest points on the head surface. Our method outperforms other methods in images taken from different angles with various expressions.