Supplementary for CLIP-Head

Pranav Manu
CVIT, IIIT-H
Hyderabad, India
pranav.m@research.iiit.ac.in

Astitva Srivastava
CVIT, IIIT-H
Hyderabad, India
astitva.srivastava@research.iiit.ac.in

Avinash Sharma
CVIT, IIIT-H
Hyderabad, India
asharma@iiit.ac.in

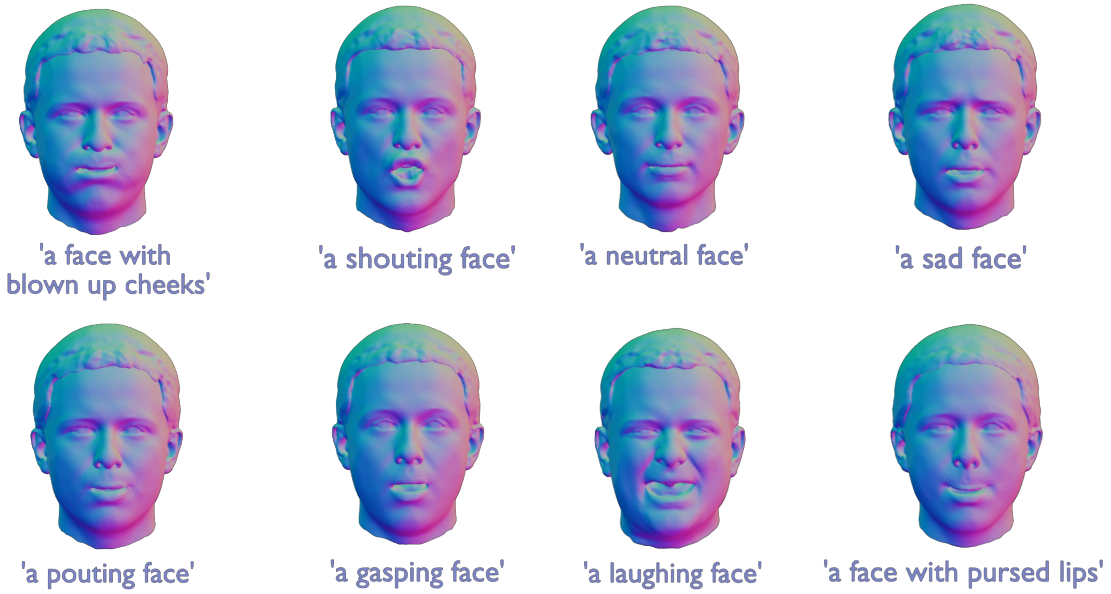

Figure 1: CLIP-Head enables text-driven generation of NPHM meshes in a variety of facial expressions.

ACM Reference Format:
Pranav Manu, Astitva Srivastava, and Avinash Sharma. 2023. Supplementary for CLIP-Head. In *SIGGRAPH Asia 2023 Technical Communications (SA Technical Communications '23), December 12–15, 2023, Sydney, NSW, Australia.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3610543.3626169

1 USER STUDY

We conduct subjective study with around 70 users to assess the quality of our generations w.r.t. existing SOTAs, ClipFace and HiFi-Face. The study was done in two settings. First one involved randomly showing generation of one of the methods and asking the user to rate on a scale of 1 to 5 with regard to whether or not it matches the prompt. Second one involved giving the output of all three methods on the same prompt and ask which one is more preferred. HiFi-face scored 1.38, ClipFace score 2.77 and CLIP-Head(ours) scored 2.85 on an average. In the second study, CLIP-Head was preferred by 41 users and ClipFace by 29 users, while none of them chose HiFi-Face.

2 IMPLEMENTATION & TRAINING DETAILS

We employ two MLPs as the proposed mapping networks to map from CLIP's latent space to NPHM's latent space. Each MLP has 8 layers with one skip connection to the 4^{th} layer and is trained for 100 epochs on 10,000 samples with a batch size of 10, using AdamW optimizer. During training, we minimize the $L2$ loss between the predicted and the sampled ground truth NPHM latent vectors. The training and the rest of the experiments are carried out on an NVIDIA RTX 4090 GPU. Our implementation for aligned-UV parameterization utilizes ***libigl*** library for computing the harmonic weights.

3 TEXTURE MISALIGNMENT AND RECTIFICATION

We prepare high-resolution aligned UV texture maps (see Figure 2) for training the Texture Synthesis module. However, the input resolution $ControlNet_{uv}$ is currently limited to 512×512. Additionally, to allow for more diversity in the generation, we keep Control strength less than 1. Due to these two factors, sometimes the details (boundary and edges) in the generated UV texture map might not be perfectly aligned with the input UV normal map. This misalignment

SA Technical Communications '23, December 12–15, 2023, Sydney, NSW, Australia

ACM ISBN 979-8-4007-0314-0/23/12…$15.00
https://doi.org/10.1145/3610543.3626169

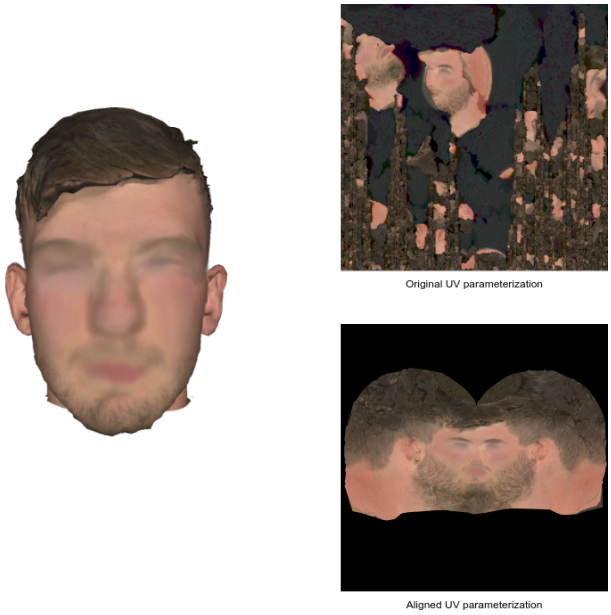

Figure 2: Re-parametrization of a sample mesh from NPHM dataset using our proposed technique to get aligned UV map with textures projected from original texture map, used for training our *Texture Synthesis* module ($ControlNet_{uv}$). Note that facial details have been blurred to protect identity.

is more evident near the facial features (eyes, nose and mouth). To rectify this misalignment, we propose a two-stage approach. First, we predict a UV texture map (possibly misaligned), conditioned on the complete UV normal map. Second, we generate a partial UV texture map from a *cropped* UV normal map (zoomed-in on the facial features), which yields a more accurate alignment, but it lacks texture information for the remaining regions (hair, ears, etc. not included). We then estimate facial landmarks using Mediapipe [Lugaresi et al. 2019] on both of the generated texture maps (complete and partial). In order to align the initially generated complete texture map, we perform TPS-based warping [Berg and Malik 2007] to deform the pixels using the facial landmarks of the initial texture map as the key points, towards the facial landmarks on the partial yet aligned texture map (see Figure 3). This rectification yields a better-aligned UV texture map with sufficient diversity.

4 EVALUATION METRICS

Here, we describe the evaluation criteria we used for evaluating two major components of our pipeline:

- **For Texture Synthesis:** We use CLIP score to evaluate the texture synthesis results. CLIP Score is defined as the cosine similarity between the CLIP-encoding of the text-prompt and CLIP-encoding of the generated image using pre-trained CLIP encoders. This metric gauges the extent to which the generated image encapsulates the essence of the text prompt. Following the approach in [Shivangi Aneja 2023], we render the textured mesh and evaluate the CLIP score between the rendered image and input text-prompt. We use the text-prompts defined by the authors of [Shivangi Aneja 2023] for evaluation. — ViT-H/14, ViT-L/14 & ViT-B/16 and demonstrate superior performance compared to SOTA methods.

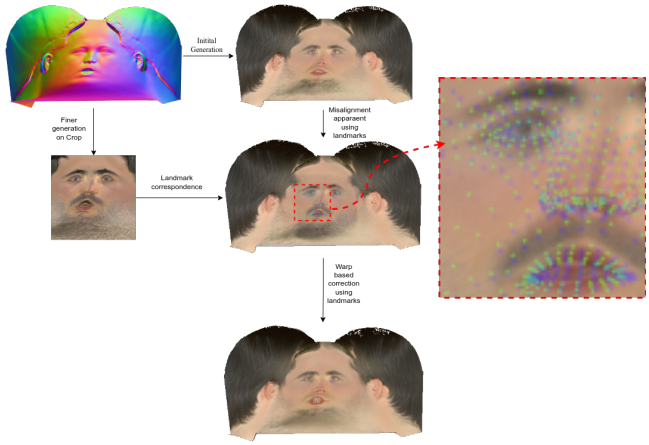

Figure 3: Texture warping for correcting the misalignment between the input UV normal map and generated UV texture map, mainly in the facial region (eyes, nose, and mouth).

- **For Aligned UV Parametrization:** We use Quasi Conformal Error (QCE) [Sander et al. 2001] and Area Scaling Error (ASE) [Sawhney and Crane 2017] for evaluation of the UV distortions while parametrizing the meshes using the proposed technique. QCE measures the angular distortion based on the ratio of the singular values of each face mapping. The ideal QCE value is 1, and a higher value implies distortion. ASE measures the scale factor of the mapped faces. Negative ASE values imply shrinkage; positives imply increase, and zero implies no area distortion in mapping. The estimated median QCE for our parametrization technique comes out to be **4.311**, whereas the median QCE of the original UV parametrization comes to be **5.143**, leading to minimal distortion. The median ASE comes out to be $-$**1.22**, whereas, for the original default parametrization, it is $-$**0.068**. This is certainly expected, as some area scaling is expected for the cost of getting aligned UV parametrization for texture synthesis. We also visualize the QCE on the test meshes from NPHM dataset in Figure 4.

5 ABLATION STUDY

With enough paired data generated, we propose to train our identity mapping network MLP_{id} for a specific CLIP embedding ψ, aiming to minimize the $L2$ loss between predicted and actual z_{id}. To enhance learning, we try to incorporate a cyclic loss L_{cyclic}, similar to [Menghua Wu 2023], introducing a backward MLP_{ψ} that takes MLP_{id}'s prediction (i.e., predicted z_{id}) to predict the original ψ. While cyclic constraints can aid network learning, we contend their benefit is limited in cases with a known intermediate space, like CLIP in our case. Deviations in backward MLP learning can adversely affect forward MLP, even when forward MLP is optimally

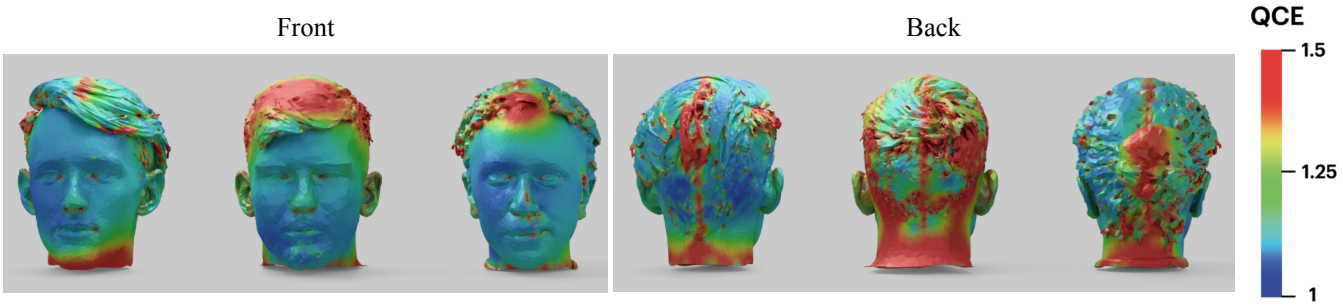

Figure 4: Visualization of QCE on NPHM dataset meshes (test samples) after performing Aligned UV parametrization. Note that most of the error is around the hair and back region (near the seam) and the error on the facial region is very low.

Loss Type	MSE (MLP_{id}) ↓	MSE (MLP_{exp}) ↓
$L2$	8.33 e-05	6.51 e-04
$L2 + L_{cyclic}$	8.74 e-05	6.87 e-04

Table 1: Ablation on cyclic loss for training mapping networks.

trained. We observe the same issue while training our expression mapping network MLP_{exp} as well. This is evident from Table 1, where the MSE (Mean Square Error) on random test samples is certainly higher when L_{cyclic} is included.

6 LIMITATIONS & FUTURE WORK

One can observe seams on the back of the head due to aligned UV parametrization; some false details on hair, due to lack of diverse hairstyles in the NPHM dataset. In future, we would like to explore disentangling the hair features as well.

REFERENCES

Alexander Berg and Jitendra Malik. 2007. *Shape Matching and Object Recognition.* Vol. 4170. 483–507. https://doi.org/10.1007/11957959_25

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. 2019. MediaPipe: A Framework for Building Perception Pipelines. *ArXiv* (2019).

Linjia Huang Yiyu Zhuang Yuanxun Lu Xun Cao Menghua Wu, Hao ZhuB. 2023. High-fidelity 3D Face Generation from Natural Language Descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*

Angela Dai Matthias Niessner Shivangi Aneja, Justus Thies. 2023. ClipFace: Text-Guided Editing of Textured 3D Morphable Models. In *ACM SIGGRAPH 2023 Conference Proceedings* (Los Angeles, CA, USA) *(SIGGRAPH '23).*