

Implicit Neural Head Synthesis via Controllable Local Deformation Fields

- Supplementary Material -

Chuhan Chen^{1*}

Matthew O’Toole¹

Gaurav Bharaj²

Pablo Garrido²

¹Carnegie Mellon University ²Flawless AI

1. Implementation Details

TensorRF We use TensorRF [2] for the canonical space of our pipeline with the following architecture changes, we use 1) an appearance latent code into the MLP decoder to account for appearance inconsistencies, and 2) RELU activation to threshold volume densities instead of Softplus to allow sharper reconstruction. To combine the training of TensorRF and deformation fields, we freeze the deformation field and pre-train TensorRF for 14k iterations. During pre-training, we grow the voxel grid from 128^3 to the maximum resolution 300^3 (as in the original paper) for *Subject1*, *Subject2*, and *Subject4* and 200^3 for *Subject3*. We also prune voxels with density smaller than $1e^{-4}$.

Deformation Field The architecture of the deformation field is shown in Fig. 7. Each local deformation field consists of a 3-layer MLP. Each layer consists of 40 neurons, followed by a Leaky RELU. The input to the input layer of each local field MLP is concatenation of a global expression code and jaw pose masked with the Attention Mask, global deformation latent code, and head and neck pose.

RigNeRF* We modify the original RigNerf [1] architecture for the monocular setting and refer to it as RigNeRF*, where the head pose from the tracker is transformed into a camera extrinsic matrix, as if the head remains static and the camera moves. Hence, for each frame, we query the mesh deformed by expression only, instead of both pose, and expression as in the original paper. We use the same network architecture, and parameters for the deformation field, as well as the same latent code dimensions. Finally, we train each sequence for 1M iterations with ray batch-size of 1550 instead of the 100k epochs indicated in the original paper, as it is a less complex task for the deformation field to learn the deformation due to expressions only rather than both pose and expression.

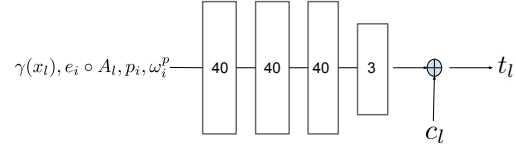


Figure 7. Architecture diagram of a single local field D_l .

		Neck			Head			Jaw		
<i>Subject1</i>	MEAN	6.09	0.314	-1.14	1.53	-6.44	-0.105	7.37	-0.0643	0.258
	STD	6.75	4.21	6.02	10.1	15.10	6.63	1.89	2.26	3.24
<i>Subject2</i>	MEAN	1.27	-0.432	0.143	3.63	-1.69	-2.67	3.94	-1.02	-1.26
	STD	7.64	7.42	6.77	9.66	32.3	7.60	2.44	3.16	6.51
<i>Subject3</i>	MEAN	-2.75	0.236	-0.794	-1.35	4.02	2.55	6.60	0.781	-0.981
	STD	2.41	1.20	3.21	3.84	4.76	3.66	1.93	0.936	2.34
<i>Subject4</i>	MEAN	3.11	-1.83	-2.87	-5.53	-2.07	2.65	4.54	-0.491	0.714
	STD	4.70	5.99	2.98	6.33	16.8	4.19	2.56	2.18	4.68

Table 3. Pose distribution of 4 sequences in the order of yaw, pitch, roll in degrees.

2. Dataset Analysis

Our video dataset consists of four subjects, as shown in Fig. 4 (main paper). *Subject1* (1st column) and *Subject2* (4th column) are subjects captured indoors with a 4K phone camera. *Subject3* (3rd column) is ex-president Obama addressing a commencement speech, and is a segment of the HD video downloaded from YouTube.¹ *Subject4* (2nd column) is a female subject from IM Avatar benchmark dataset.² The first three datasets show unscripted natural expressions with varying head poses, while the latter is split into a speech video and another video with difficult expressions and poses, as described in [5]. Tab. 3 shows the neck, head, and jaw pose distribution of the four monocular sequences. Please refer to the supplementary video for detailed visualization of the poses as well as the range of expressions.

3. Additional Results

Test-time latent code optimization (deformation and appearance) helps adjust pose inaccuracies, which is a cur-

¹Commencement speech of class 2020: <https://youtu.be/NGEvASSaPyg>

²https://dataset.ait.ethz.ch/downloads/IMavatar_data/data/yufeng.zip We use sequence MVL1810 and MVL1814 as training set and MVL1812 as test set

*Work was done while interning at Flawless AI.

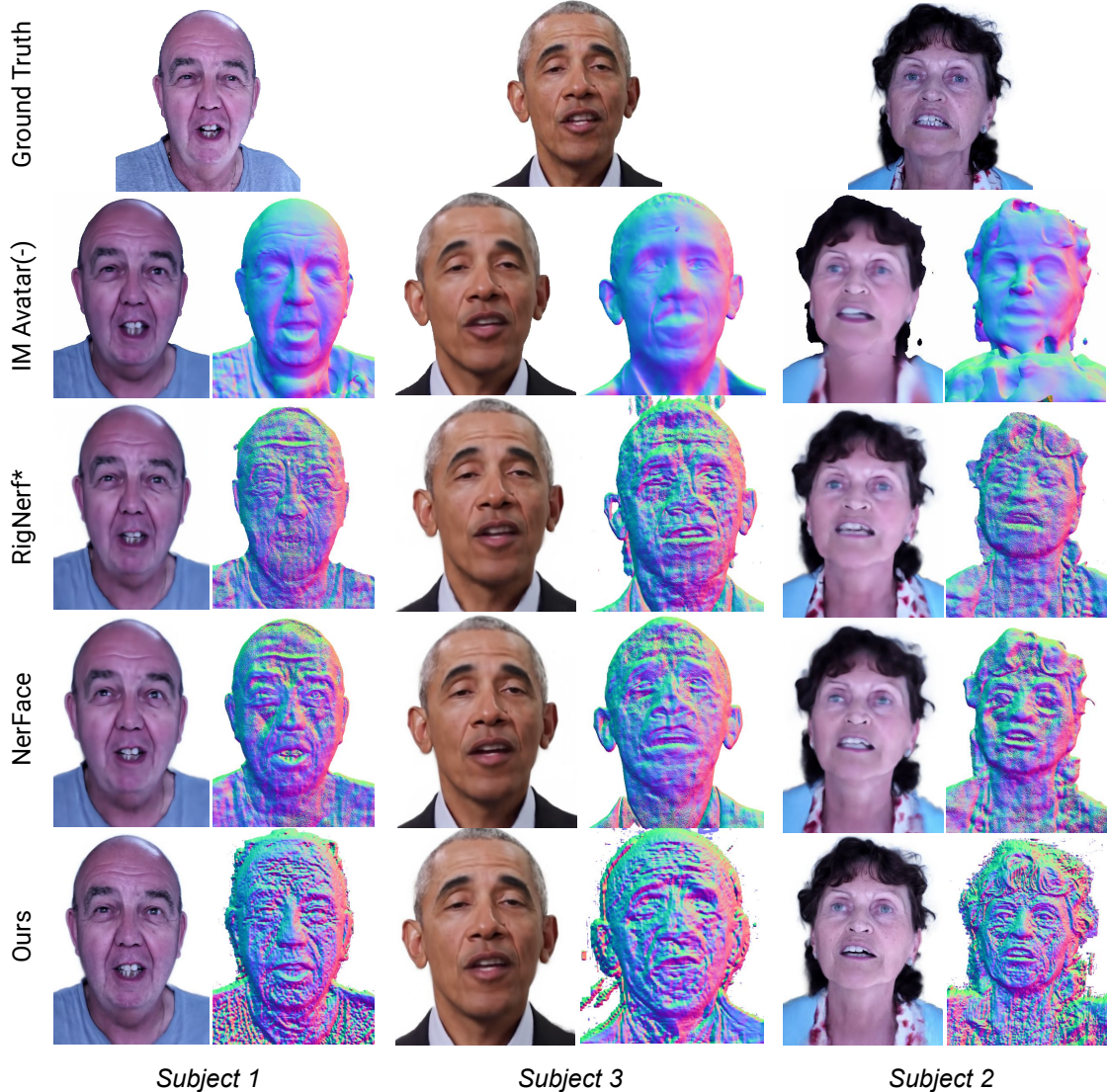


Figure 8. **Qualitative comparisons of geometry with state-of-the-art on test data without tuning of latent code during test time.** Top to bottom: GT, IM Avatar(-), RigNerF*, NerFace, and our approach. Here, the images are rendered with the latent codes of the first training frame of each sequence. Note that our approach produces significantly richer geometric details, as observed in the normal maps. Besides, the rendered images generated by our method faithfully reflect the pose, expression, and appearance of the ground truth images.

rent limitation of our approach. However, our model excels at producing high-quality reconstruction, even without per-frame latent code optimization. Fig. 8 compares the quality of the normals obtained by our methods and the different baselines on *Subject1*, *Subject2*, and *Subject3*. Note that no image detail is lost when compared to Fig. 4 (main paper). Besides, our method produces crisper results than baseline approaches, as shown in the reconstructed normal maps.

Tab. 4 shows additional metrics for Ours[†] and RigNerF*[†], both using no optimized latent code. Ours[†] suffers from global pose misalignment (several pixels off)

without optimized latent code, impacting per-pixel image metrics (PSNR and ℓ_1). Local facial structures are still well preserved as demonstrated by consistently lower scores (second best) on LPIPS for which we still achieve state-of-the-art performance.

4. Extended Analysis

Modeling Local Deformation Fields In our implementation, we model local deformation fields around a subset of facial landmarks using DECA’s landmark definition [3]. Specifically, we experimented with 5 and 34 keypoint lo-

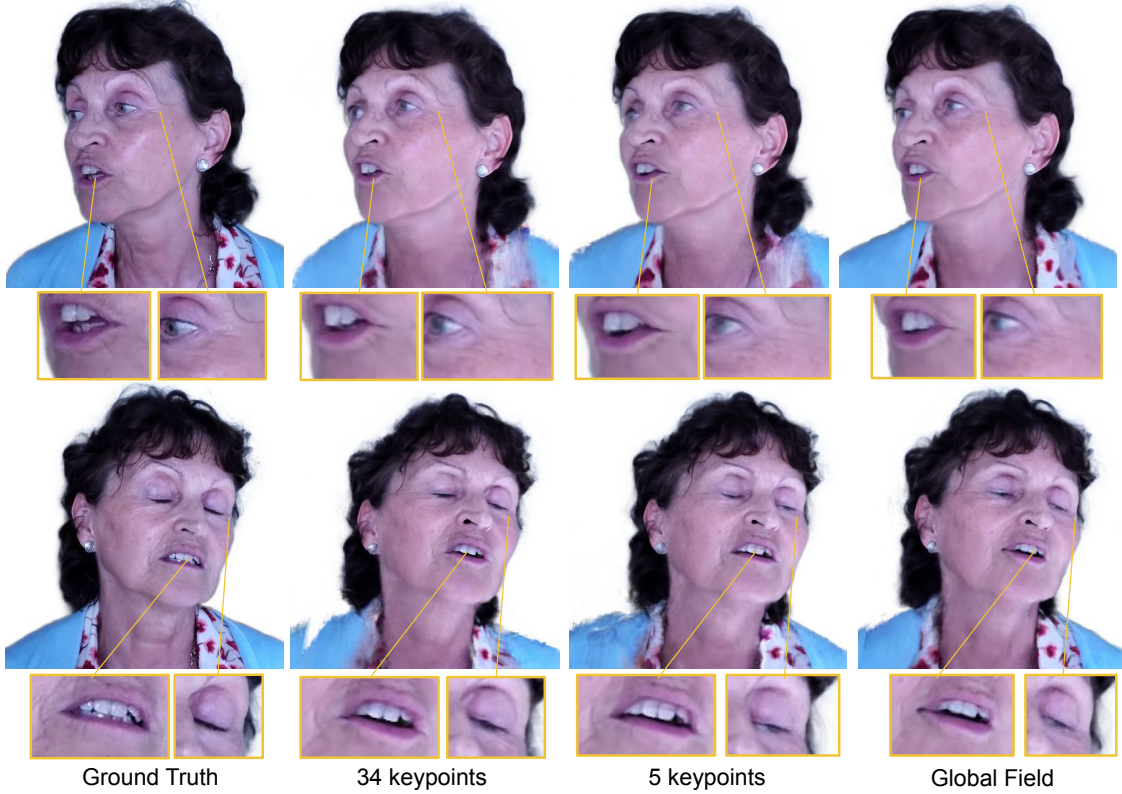


Figure 9. **Ablation on the number of keypoints used by our method.** From left to right: ground truth, 34 keypoints, 5 keypoints, and global field (*i.e.*, no local decomposition). Note that the network size for each experiment is adjusted such that the total number of parameters is roughly the same. Our method produces better visual results with 34 keypoints.

	Subject1				Subject2			
	$\ell_1 \downarrow$	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	$\ell_1 \downarrow$	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow
Nerface	0.058	0.903	0.105	21.457	0.077	0.889	0.121	18.252
IM Avatar(-)	0.068	0.901	0.113	20.502	0.093	0.877	0.157	14.960
RigNeRF*	0.055	0.904	0.095	22.324	0.072	0.884	0.120	18.922
RigNeRF*†	0.0767	0.876	0.108	18.829	0.0814	0.881	0.131	17.598
Ours†	0.0773	0.895	0.0611	19.681	0.0873	0.876	0.0795	16.917
Ours	0.054	0.929	0.051	23.508	0.062	0.917	0.0576	20.4375
	Subject3				Subject4			
	$\ell_1 \downarrow$	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	$\ell_1 \downarrow$	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow
Nerface	0.047	0.894	0.055	22.220	0.077	0.818	0.085	17.910
IM Avatar(-)	0.043	0.900	0.078	23.218	0.063	0.870	0.069	19.215
RigNeRF*	0.035	0.910	0.052	24.634	0.065	0.844	0.063	19.253
RigNeRF*†	0.03430	0.898	0.0526	23.109	0.118	0.727	0.129	14.677
Ours†	0.0335	0.925	0.0387	24.552	0.0745	0.789	0.0453	16.834
Ours	0.0206	0.971	0.0265	30.854	0.081	0.830	0.062	19.085

Table 4. Extended quantitative comparisons for Tab. 1. Ours† and RigNeRF*† are run without optimized per-frame deformation latent code. **Bold black** is best result; **blue** is second best.

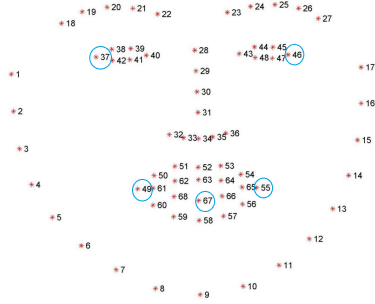
cations, as shown in Fig. 10. The former representation is similar to [4], though we change the tip of the nose with the midpoint of the lower lip to better model jaw deformations. In the latter, denser representation, we exploit the semantics of the facial landmark definition and center local fields around every other landmark.

Fig. 9 compares the effect of using different numbers of keypoints. Note that our method reconstructs sharper de-

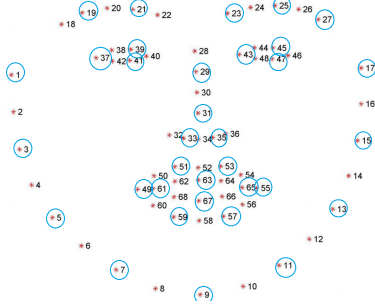
tails and more accurate facial features when the deformation field is decomposed with 34 keypoints. Overall local decomposition with both 34 and 5 keypoints results in better reconstruction of details than that of the global field. The latter tends to over smooth the reconstructed surface and produce less accurate facial deformations than multiple local deformation fields.

References

- [1] ShahRukh Athar, Zexiang Xu, Kalyan Sunkavalli, Eli Shechtman, and Zhixin Shu. Rignerf: Fully controllable neural 3d portraits. *CoRR*, abs/2206.06481, 2022. 1
- [2] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. *CoRR*, abs/2203.09517, 2022. 1
- [3] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM TOG*, 40(4):88:1–88:13, 2021. 2
- [4] Mingwu Zheng, Hongyu Yang, Di Huang, and Liming Chen. Imface: A nonlinear 3d morphable face model with implicit neural representations. *CoRR*, abs/2203.14510, 2022. 3



(a) locations of 5 keypoints



(b) locations of 34 keypoints

Figure 10. keypoint locations for experiments

- [5] Yufeng Zheng, Victoria Fernández Abrevaya, Xu Chen, Marcel C. Bühler, Michael J. Black, and Otmar Hilliges. I M avatar: Implicit morphable head avatars from videos. *CoRR*, abs/2112.07471, 2021. 1