

Supplementary Material - SplatFace: Gaussian Splat Face Reconstruction Leveraging an Optimizable Surface

Jiahao Luo¹ Jing Liu² James Davis¹
¹University of California, Santa Cruz ²ByteDance Inc.
{jluo53, davisje}@ucsc.edu

1. Videos

We includes 6 samples with various identities in the supplemental. It shows a comparison with other Gaussian splatting methods on the FaceScape dataset with 4-image inputs. We show an extrapolation from center to approximately 30 degree to the top left or top right. Figure 2 shows a qualitative comparison from two examples. 3DGS and Mip-Splatting generates floaters and spikes when the angle is large. FSGS generates overly smooth results and has lighting problems. Our method produces the most visually pleasing results with the fewest artifacts.

2. Comparison with GaussianAvatar

GaussianAvatar [3] is designed for animation (rather than static reconstruction) and with far more views (16 views rather than 3-5 views). It relies on pre-determined 3DMM parameters and only finetunes expression and pose. However, GaussianAvatar attaches Gaussians on the mesh triangles which is similar to our proposed non-rigid alignment. Therefore, we believe it’s interesting to compare. GaussianAvatar regularizes position with a point-to-triangle distance that only uses the splat center, and regularizes covariance with a simple preference for smaller scale splats. This simpler loss works when many views are available, but breaks down with fewer views. Fig 1 provides a direct comparison to their provided implementation using NeRSemble data from their paper, restricted to 4 views. Our method produces results with fewer artifacts.

3. Comparison with NeRF

Due to limited space in the main text, we showed only a comparison with other Gaussian Splatting methods. The NeRF family of techniques is closely related, so we provide here a comparison with the state-of-the-art 3D human face NeRF method, DINER [2]. This method was chosen because it’s the most recent openly available method and specifically works well on 3D faces with only 4-image inputs. The authors reports higher performance than other

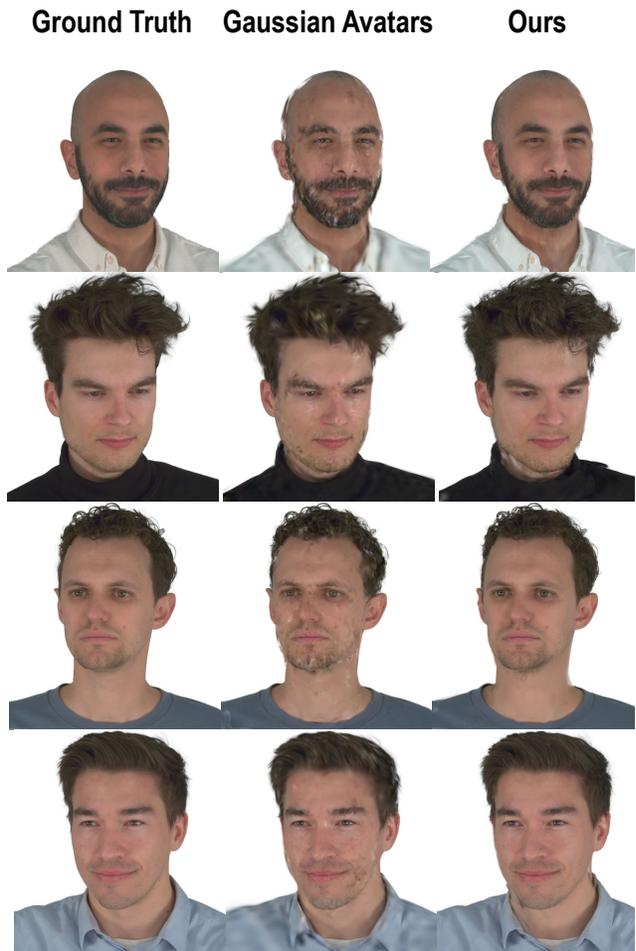


Figure 1. A qualitative comparison of novel view synthesis to Gaussian Avatars with 4-view input.

few-view NeRF methods [1,5]. Figure 3 shows a qualitative comparison between our method and DINER. Our method produce results with more high-frequency details and fewer artifacts.

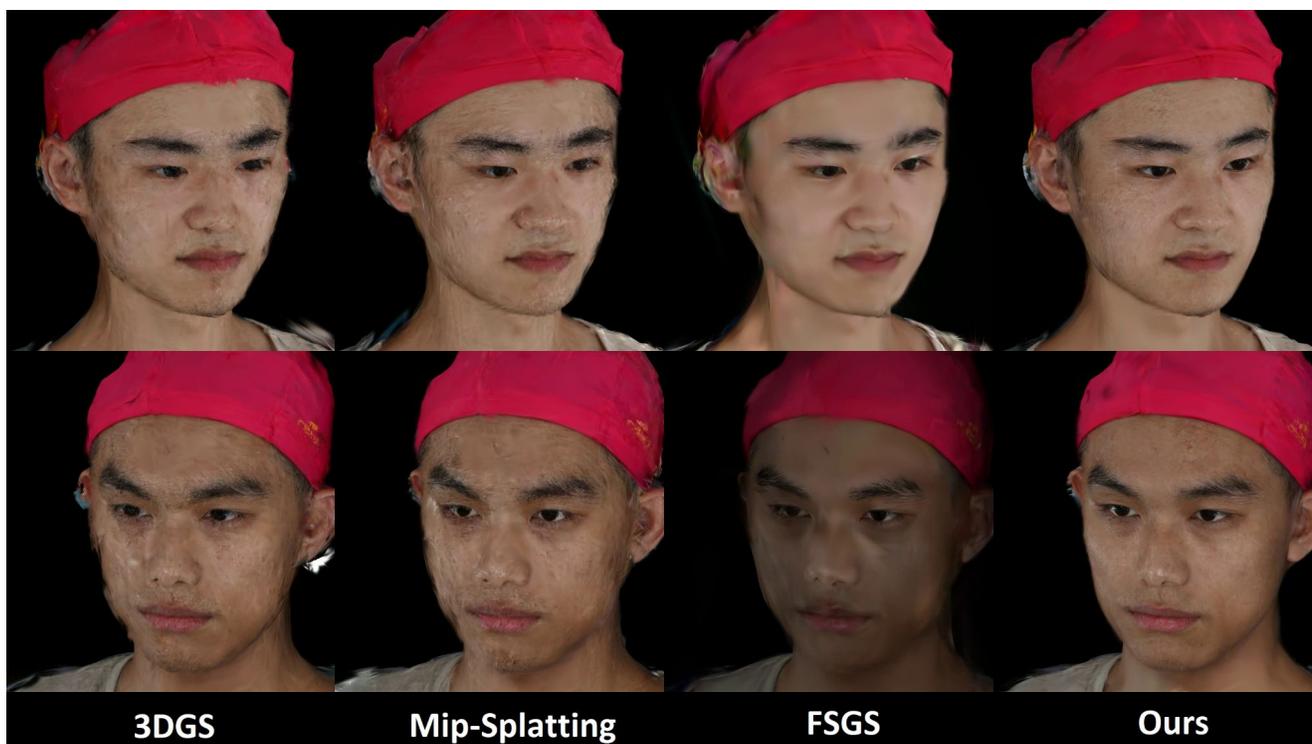


Figure 2. Sample video frame

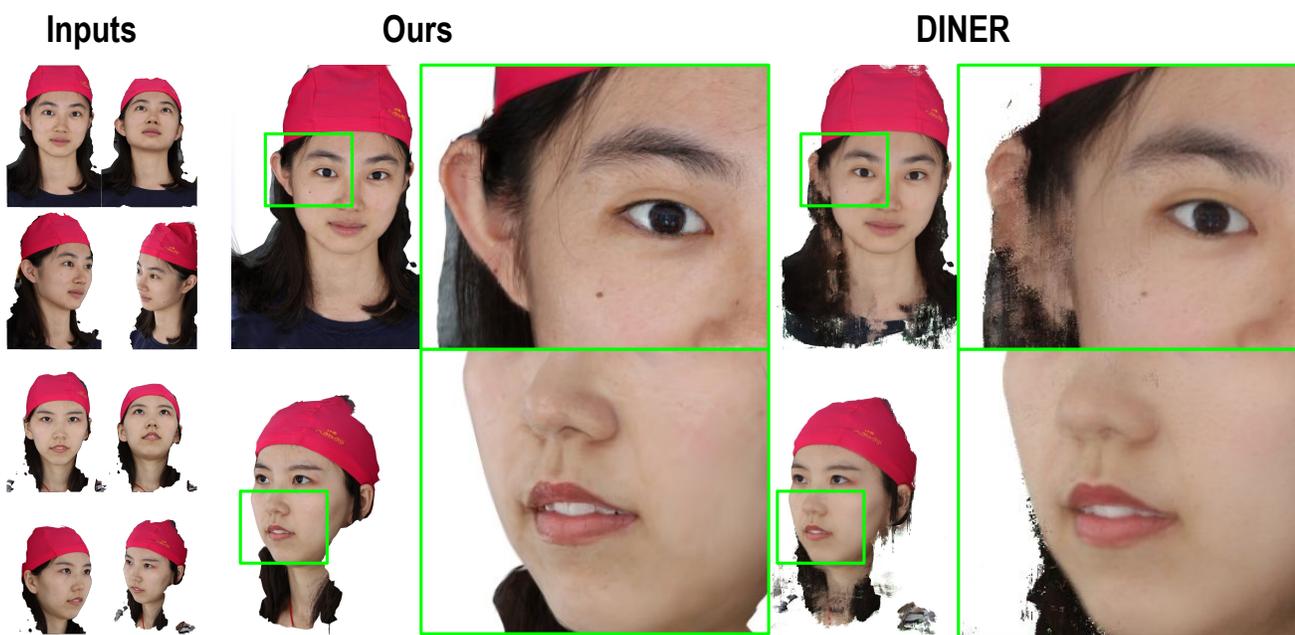


Figure 3. Qualitative comparison between our method and DINER on FaceScape dataset. The selected test view is close to the training views. Our method produce results with more high-frequency details and fewer artifacts.

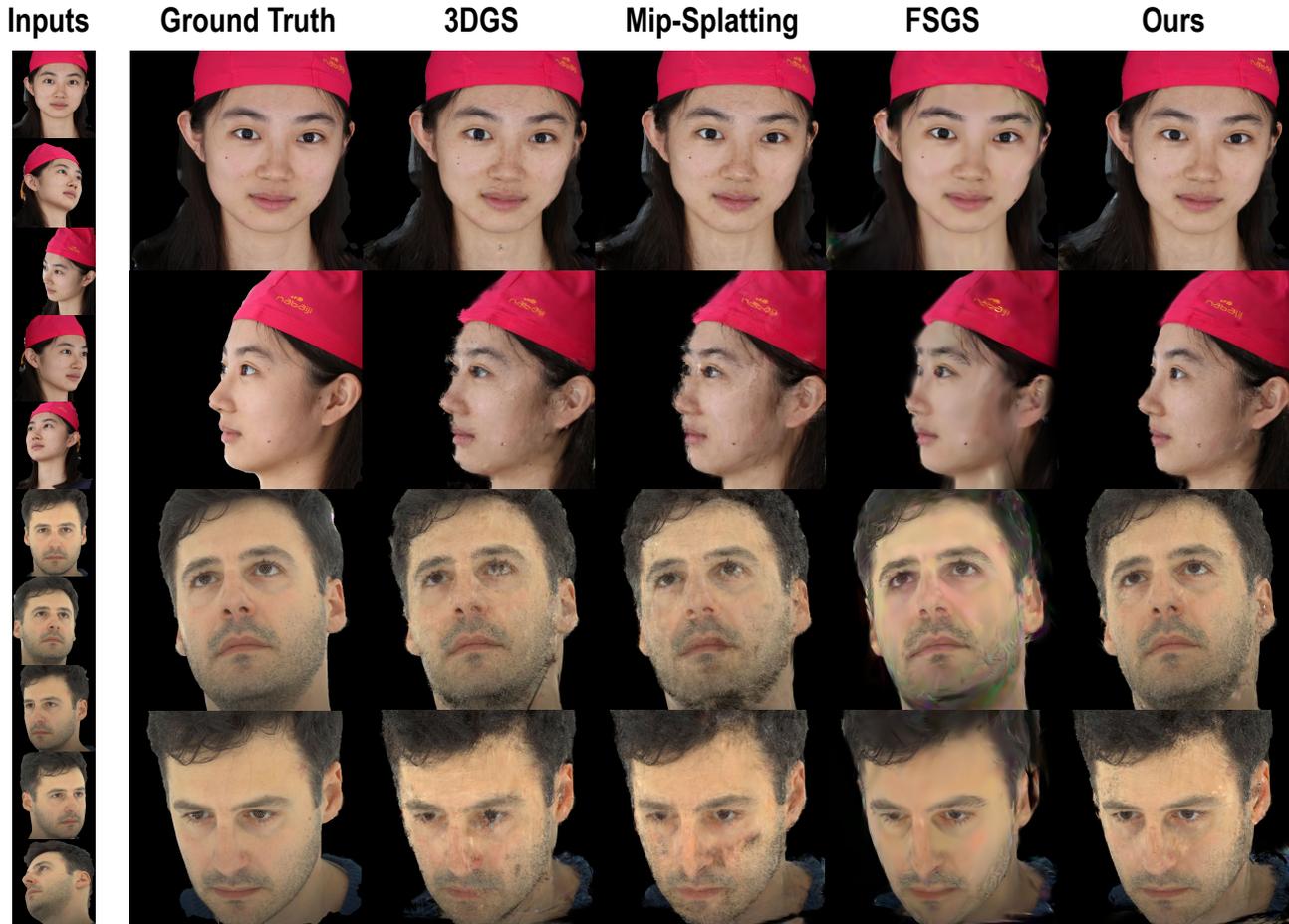


Figure 4. Qualitative comparison on novel view synthesis with different viewpoints. For each individual, a test view that is close to the training views is shown in the top row, and a test view further from the training images is shown in the bottom row. Our method produce results with fewer artifacts than the comparison methods.

4. Comparison with different viewpoints

Figure 4 shows an evaluation on both a near and far test viewpoint. In the first row of each individual, we show a test view near to one of the training views. 3DGS and Mip-splatting tend to produce noisy results, whereas FSGS often yields overly smoothed outcomes. In contrast, our method succeeds in capturing high-frequency details with minimal artifacts. In the second row for each test subject, we show an example which is far from the training views. Extrapolation of viewpoint far away from training views is very challenging and we do not expect perfect results from any method. 3DGS and Mip-Splatting produce noisy results and exhibit floating splats due to the lack of geometric constraints. These floating splats are most visible in profile views since they lie obviously away from the face. FSGS results in mismatched colors and poor geometry. While our method also contains artifacts, it yields the most visually appealing novel view synthesis.

5. Comparison with FlashAvatar

FlashAvatar [4] is a method that uses a monocular video sequence during optimization. The results in that paper show animation from the same viewpoint as the training video. It is perhaps unfair to analyze the method on a scenario it was not designed for. Nevertheless, in order to see if monocular input generalizes to novel viewpoints we include an example video while rotating the viewpoint. Figure 5 shows one frame of this video. Notice that rendering quality is significantly degraded when the viewpoint is changed.

6. Sensitivity Analysis of N

When sampling each Gaussian Splat to compute the SplatToSurface loss, the number of samples per iteration, N , has only a mild effect on visual quality. The samples are randomly chosen in each iteration, so each Gaussian is sampled thousands of times over the course of optimization.

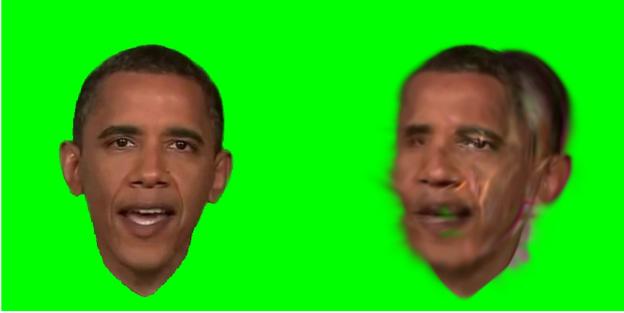


Figure 5. Monocular methods like FlashAvatar are not designed for novel view synthesis and thus perform poorly when rendering viewpoint is changed.

N	1	2	4	6
time	8.0	11.3	18.2	25.2
L1	0.0247	0.0233	0.0239	0.0228
SSIM	0.8541	0.8556	0.8489	0.8544
PSNR	26.44	26.58	26.75	27.41
LPIPS	0.1266	0.1193	0.1184	0.1217

Figure 6. A sensitivity analysis of the effect of modifying the number of samples, N , used when sampling a Gaussian Splat reveals that the number of samples has only a mild effect on visual quality.

Lambda	100	500	1000	1500	2000	10000
L1	0.0245	0.0247	0.0233	0.0231	0.0250	0.0271
SSIM	0.8537	0.8443	0.8556	0.8577	0.8556	0.8436
PSNR	25.82	26.12	26.58	26.03	25.88	25.45
LPIPS	0.1377	0.1340	0.1193	0.1229	0.1340	0.1696

Figure 7. A sensitivity analysis of the effect λ_{s2s} , the parameter controlling the relative contribution of different loss terms. A relatively wide range of values produce adequate performance.

tion. Figure 6 shows a comparison on $N=1,2,4,6$, and lists optimization time in minutes, as well as four image quality metrics (L1, SSIM, PSNR, LPIPS). Notice that increased samples has little effect. All experiments in this paper were conducted with $N=2$.

7. Sensitivity Analysis of λ_{s2s}

The relative weight of multiple loss terms in our method are controlled by the parameter λ_{s2s} . Figure 7 shows multiple settings of this parameter and the effect on several measures of image quality. A relatively wide range of values are acceptable. All experiments in this paper use $\lambda=1000$.

References

- [1] Marko Mihajlovic, Aayush Bansal, Michael Zollhoefer, Siyu Tang, and Shunsuke Saito. Keypointnerf: Generalizing image-based volumetric avatars using relative spatial encoding of keypoints. In *European conference on computer vision*, pages 179–197. Springer, 2022. 1
- [2] Malte Prinzler, Otmar Hilliges, and Justus Thies. Diner: Depth-aware image-based neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12449–12459, 2023. 1
- [3] Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. *arXiv preprint arXiv:2312.02069*, 2023. 1
- [4] Jun Xiang, Xuan Gao, Yudong Guo, and Juyong Zhang. Flashavatar: High-fidelity head avatar with efficient gaussian embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1802–1812, 2024. 3
- [5] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2021. 1