VoluMe – Authentic 3D Video Calls from Live Gaussian Splat Prediction
Supplementary Material

Martin de La Gorce Charlie Hewitt Tibor Takács Robert Gerdisch Zafiirah Hosenie
Givi Meishvili Marek Kowalski Thomas J. Cashman Antonio Criminisi

Microsoft, Cambridge, UK

1. Evaluation Details

For our evaluations on Cafca and Ava-256 we follow the process of Lyu et al. [2]. For Cafca the renders are 512×512 pixels and include perfect camera parameter annotations. We use the same subset of the data as Lyu et al. [2], which includes 40 identities with between 9 and 19 views per identity. For Ava-256 the images are 667×1024 pixels and well-calibrated camera annotations are provided. Again, we use the same subset of the data as Lyu et al. [2], including 10 identities with 10 views per identity. The background of the images is removed using the ground-truth masks and replaced by white pixels before we provide the input image to the reconstruction process. The ground-truth camera parameters are then used to render several novel views. Lyu et al. [2] align these cameras to their coordinate system while we choose to align our 3D result to the coordinate system of each dataset. The rendered views and ground-truth images are then aligned using 2D landmark detection to best evaluate the visual quality of the reconstruction, rather than the quality of the alignment. Note that we only evaluate using the 2D-aligned protocol for Cafca as the unaligned protocol uses all 30 views per subject, these include many views of the rear of the head which we do not target with our method. For Ava-256 the results are cropped and resized to 512×512 before calculating metrics. The metrics code is not available so we have re-implemented to the best of our ability with the help of the authors, we do however still expect there to still be discrepancies in alignment/cropping which can have a large impact on the metrics.

2. Additional Results

Fig. 1 provides additional qualitative results of our method. Fig. 2 shows further results with geometry visualization. To create these visualizations we render the depth image for Gaussians with slightly larger scale using the Gaussian Splats renderer set up to take the transparencies into account in the alpha-composition. We then calculate the surface normals from this, given the camera intrinsics. We then compute an approximate glossy render as a linear combination of the z-component of the normal and its square. Video results can be found at https://aka.ms/VoluMe.

3. Ablation Experiments

To validate the impact of the various elements of our approach we conduct a series of ablation experiments with results given in Tab. 1.

We include a baseline using the simplified U-Net backbone of our method and the same training method as Szymanowicz et al. [3], i.e., where we removed the colour sampling step, the optimized channels and jitter loss, used a single Gaussian per pixel and used simple cropping with CLIFF parameters [1] instead of the homography. We kept for this baseline the re-scaling step to allow training with variable camera-to-face distances as we did not succeed training a model without this term. We provide the results of our full method for comparison against this baseline, and our method omitting each modification as listed above and described in the main paper in sequence. For these ablation experiments, the original Ava256 dataset subsets used in Lyu et al. [2] and described in Sec. 1 appeared to be too small to allow reliable numerical comparisons between the different methods. We used instead a much larger subset of the original Ava256 datasets. We used 250 subjects out of 256 original subjects and for each subject we select a single pair of consecutive frames. We removed from the original dataset 6 subjects (20220815–1656–MOP211, 20230906–0803–HBM931, 20230817–1136–DZS003, 20230918–1040–NTA876, 20210928–0843–PAK800, 20230310–1106–FCT871) that did not have one of the two cameras that we chose as input cameras (401168 and 401875). For each subject the input view was chosen as the camera with index 401168 (which was chosen as the input camera by Lyu et al. [2]) when it existed, or the camera 401875 (whose centre was closest to the world origin) otherwise. We then chose the 10 views closest to the input camera centre to compute the metrics. To compute the jitter

Figure 1. Additional qualitative results of our method.

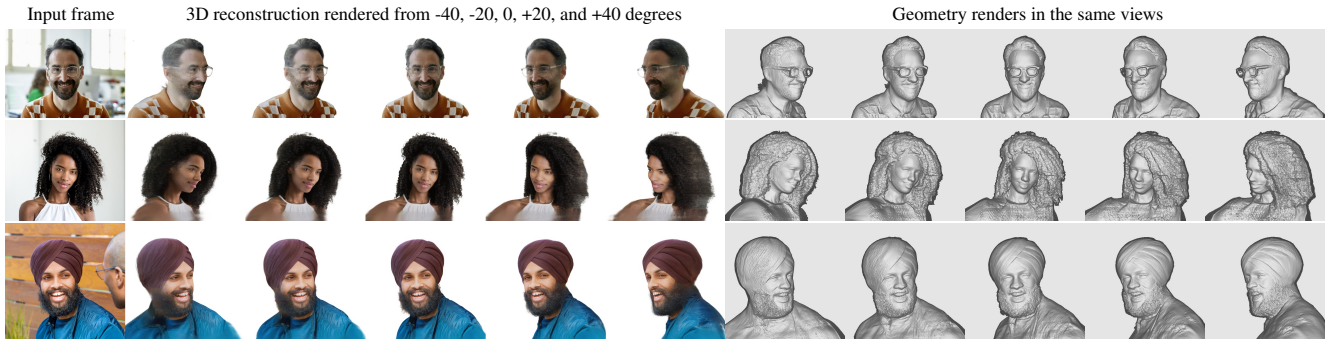

Figure 2. Additional qualitative results of our method with geometry renders.

metric, we used a pair of consecutive frames.

The main results from this ablation experiments are: (1) our best method that combines all the different technical contributions of the paper performs much better than the baseline approach across all metrics and performs best or second to best on all but one metric (DreamSim). (2) The model trained without the jitter loss has worse performance on the jitter metric than the model trained with it, which illustrates the benefit of this loss specifically to reduce jitter.

3.1. Impact of multiple Gaussians per pixel

Fig. 3 shows renders for each of the two Gaussians per pixel that our model predicts in isolation. One 'layer' of Gaussians is representing the coarse appearance and has clearly learnt a strong prior on the underlying 3D geometry, similar to the results of Szymanowicz et al. [3]. The other layer is behaving more like a monocular depth prediction model, predicting a sheet of small Gaussians which correspond to the first visible surface in the image. This second layer

Table 1. Ablation results on the Ava-256 dataset, following the protocol of Lyu et al. [2] with a larger subset of the original datasets and 100 training epochs. All results are with facial landmark alignment. Best results are **bold** and second best underlined. The method that combines all our features is first on all metrics but the jitter metric. The baseline method is the worst on all metrics but the jitter metric.

Method					PSNR ↑	SSIM ↑	LPIPS ↓	DreamSim ↓	ArcFace ↓	Jitter ↓
Two Gaussian Per Pixel	ROI Homography	Jitter Loss	Direct Sampling	Optimizable Layers						
×	×	×	×	×	18.698	0.8599	0.1587	0.0553	0.1094	0.004959
✓	✓	✓	✓	✓	**19.069**	**0.8682**	**0.1497**	**0.0468**	**0.0991**	0.004967
×	✓	✓	✓	✓	18.907	0.8676	0.1528	0.0498	0.1012	**0.004856**
✓	×	✓	✓	✓	18.901	0.8663	0.1562	0.0574	0.1018	0.005034
✓	✓	×	✓	✓	18.949	0.8658	0.1581	0.0499	0.1027	0.005020
✓	✓	✓	×	✓	19.026	0.8677	0.1512	0.0565	0.0993	0.004925
✓	✓	✓	✓	×	19.005	0.8660	0.1585	0.0508	0.1061	0.005170

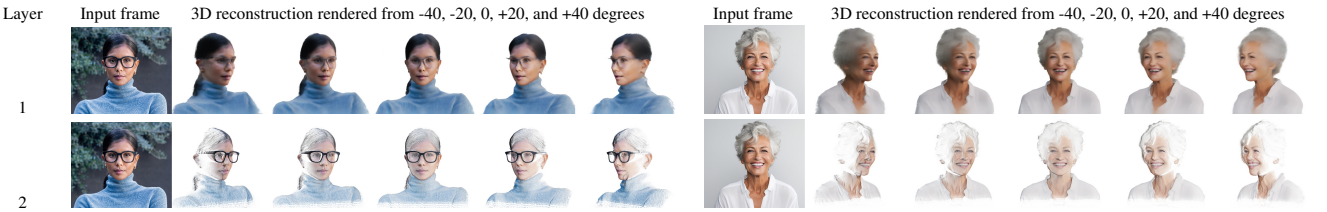

Figure 3. Results showing each 'layer' of Gaussians predicted by our method rendered separately. High frequency details and disconnected structures such as glasses are represented by one layer, and the underlying face geometry represented by the other.

Table 2. Impact of number of Gaussians per pixel. Results on the Ava-256 dataset, following the protocol of Lyu et al. [2] with a larger subset of the original datasets and 60 training epochs. All results are with facial landmark alignment.

Gaussians per Pixel	PSNR ↑	SSIM ↑	LPIPS ↓	DreamSim ↓	ArcFace ↓	Jitter ↓
1	19.07	0.8668	0.1556	0.05576	0.1114	0.004818
2	**19.15**	**0.8690**	**0.1518**	**0.04677**	**0.1003**	0.005086
3	18.99	0.8663	0.1588	0.05020	0.1073	**0.004784**
4	19.06	0.8672	0.1567	0.05220	0.1040	0.005990

is therefore able to capture high frequency details that are lacking in previous work, while the overall geometry is still well represented by the first layer. It is also interesting to note that disconnected structures such as glasses frames are captured well by the second layer, and the first layer can then infill any occluded parts on the face surface behind, based on prior knowledge learnt from the training data. The positive impact of using multiple Gaussians per pixel rather than one is also demonstrated quantitatively in Tab. 2. Using two Gaussians per pixel improves visual quality at some cost to performance. Using more than two degrades visual quality as measured by the different metrics, and is significantly more costly in terms of performance. We believe this degradation could be due to the difficulty for the network to predict more Gaussians per pixel while keeping the number of channels fixed in the last few layers of the U-Net. Increasing the number of channels in the last few layers could help 3 or 4 Gaussians per pixel work better than 2, but would also come with an increased compute cost.

References

[1] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. CLIFF: Carrying location information in full frames into human pose and shape estimation. In *Computer Vision – ECCV 2022*, pages 590–606, Cham, 2022. Springer Nature Switzerland. 1

[2] Weijie Lyu, Yi Zhou, Ming-Hsuan Yang, and Zhixin Shu. FaceLift: Single image to 3D head with view generation and gs-lrm. *arXiv preprint arXiv:2412.17812*, 2024. 1, 3

[3] Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Splatter image: Ultra-fast single-view 3D reconstruction. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 2