GaussianAvatars: Photorealistic Head Avatars with Rigged 3D Gaussians

CVPR SEATTLE, WA JUNE 17-21, 2024

Shenhan Qian[1] Tobias Kirschstein[1] Liam Schoneveld[2] Davide Davoli[3] Simon Giebenhain[1] Matthias Nießner[1]

[1]Technical University of Munich [2]Woven by Toyota [3]Toyota Motor Europe NV/SA (associated partner by contracted service)

TOYOTA woven by TOYOTA TUM

Introduction

TL;DR: We rig 3D Gaussians to a parametric mesh model (FLAME) for photorealistic head avatar creation and manipulation.

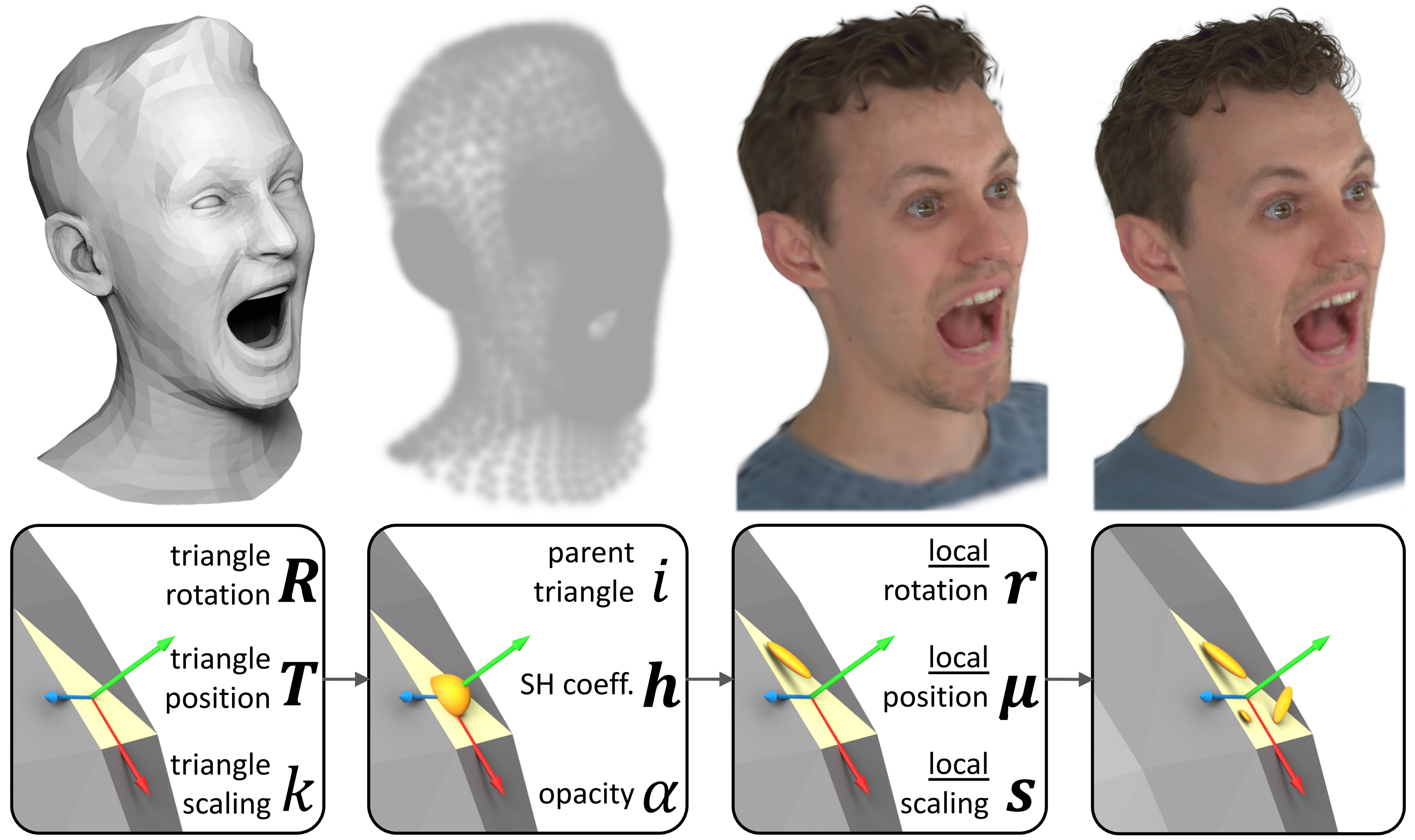
Fidelity: Using 3D Gaussian as the rendering primitive, we capture fine details like hair strands and eyelashes.

Controllability: We can manipulate Gaussian splats with FLAME as the control interface, independent of the number of splats.

Connectivity: The triangles implicitly bundle adjacent splats, enforcing smoothness on the movement of surrounding splats.

Consistency: A Gaussian splat is consistently attached to the same triangle during optimization. This enables effective finetuning of mesh parameters, drastically enhancing the reconstruction of visual details.

Pipeline



① Set up a local space for each triangle.
② Assign a Gaussian to each triangle.
③ Deform Gaussians during optimization.
④ Add Gaussians with binding inheritance.

Optimization

We optimize 3D Gaussian parameters from scratch and finetune FLAME's dynamic parameters with an RGB loss and two regularization terms.

FLAME parameters

static:
- shape $\boldsymbol{\beta}$
- vertex offset $\boldsymbol{\Delta v}$

dynamic:
- translation $\boldsymbol{t}$
- joint poses $\boldsymbol{\theta}$
- expression $\boldsymbol{\psi}$

3D Gaussian parameters

- global scaling $\boldsymbol{s}' = k\boldsymbol{s}$
- global position $\boldsymbol{\mu}' = k\boldsymbol{R}\boldsymbol{\mu} + \boldsymbol{T}$
- global rotation $\boldsymbol{r}' = \boldsymbol{R}\boldsymbol{r}$
- parent triangle i
- SH coefficient $\boldsymbol{h}$
- opacity α

loss terms

$$\mathcal{L}_{\text{rgb}} = (1-\lambda)\mathcal{L}_1 + \lambda\mathcal{L}_{\text{D-SSIM}}$$

$$\mathcal{L}_{\text{scaling}} = \left\|\max(\boldsymbol{s},\ \epsilon_{\text{scaling}})\right\|_2$$

$$\mathcal{L}_{\text{position}} = \left\|\max(\boldsymbol{\mu},\ \epsilon_{\text{position}})\right\|_2$$

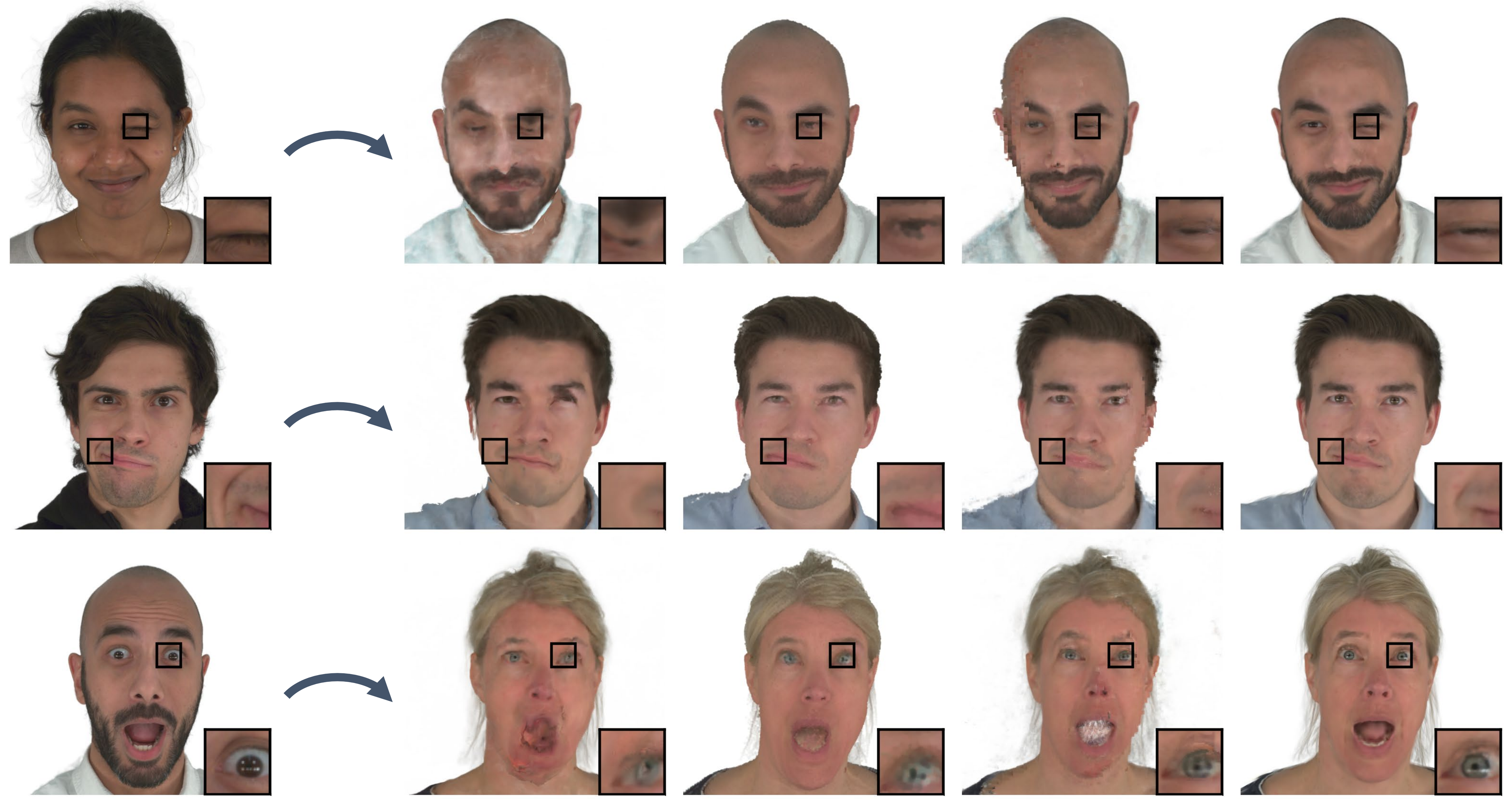
Comparison

Novel-View Synthesis: render an avatar from a left-out viewpoint.

Self-Reenactment: animate an avatar with unseen expressions and poses from the same person.

	Novel-View Synthesis			Self-Reenactment		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
AvatarMAV [47]	29.5	0.913	0.152	24.3	0.887	0.168
PointAvatar [53]	25.8	0.893	0.097	23.4	0.884	0.102
INSTA [56]	26.7	0.899	0.122	26.3	0.906	0.110
Ours	31.6	0.938	0.065	26.0	0.910	0.076

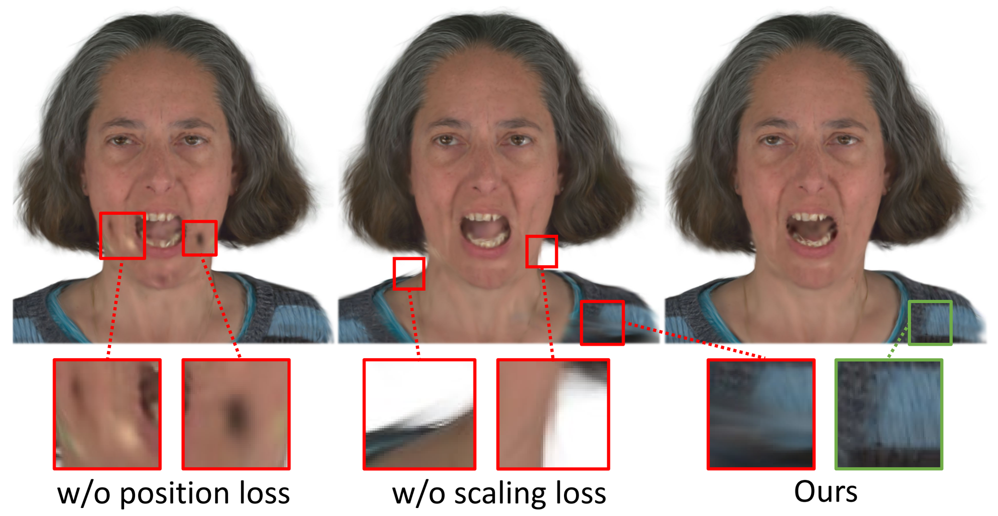
Cross-identity Reenactment: manipulate an avatar with unseen expressions and poses of another person.



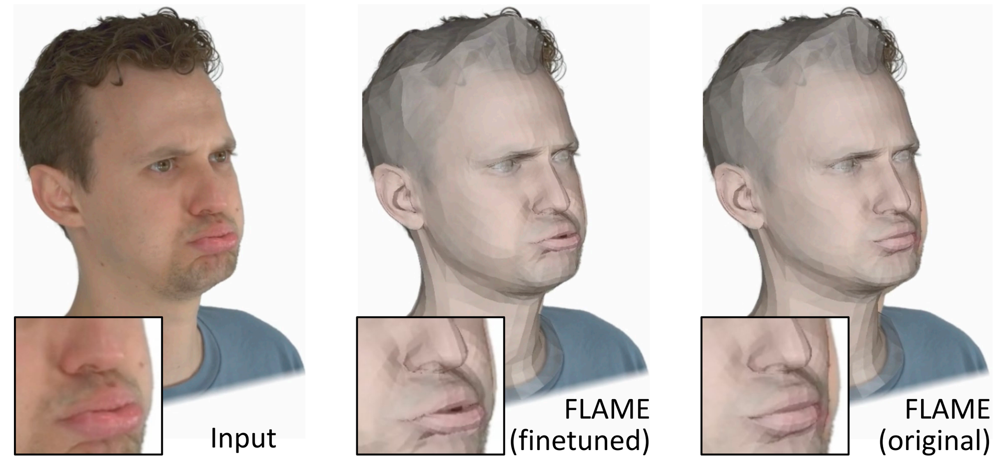
Ablation and Visualization

Regularizations prevent artifacts with novel expressions:

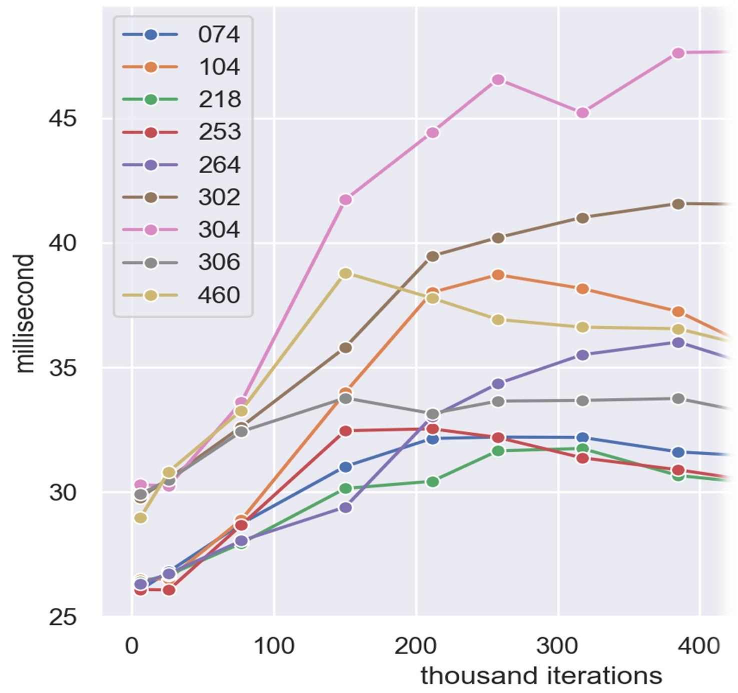
- A splat shouldn't be too far from its parent triangle.
- A splat shouldn't be too large on either axis. Smaller splats result in less blur.

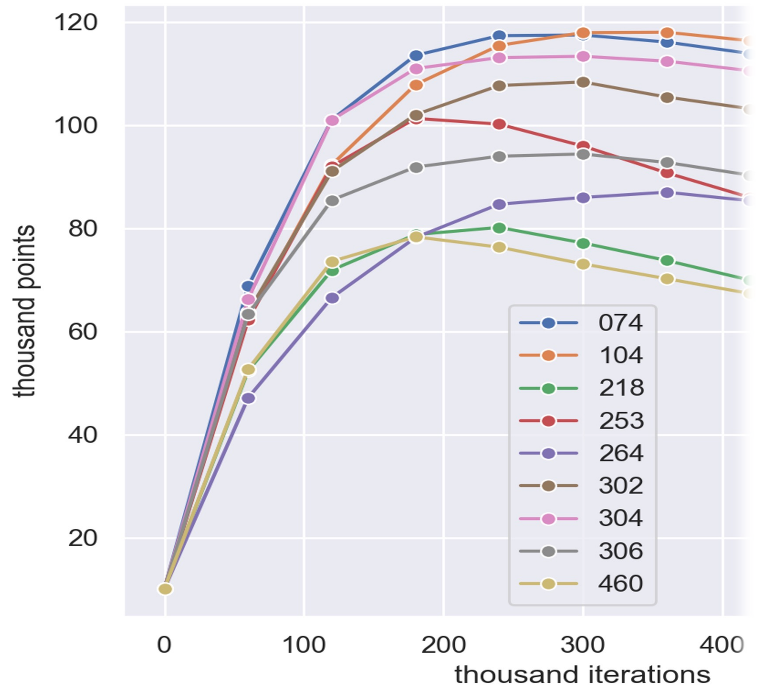


FLAME parameter finetuning:
Fine-tuning FLAME parameters leads to more accurate mesh alignment to the input image. The movement of cheeks and lips is better captured with FLAME fine-tuning.

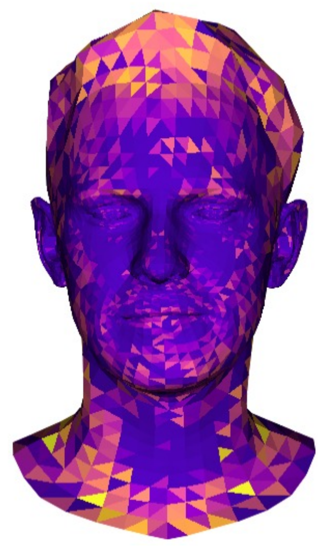


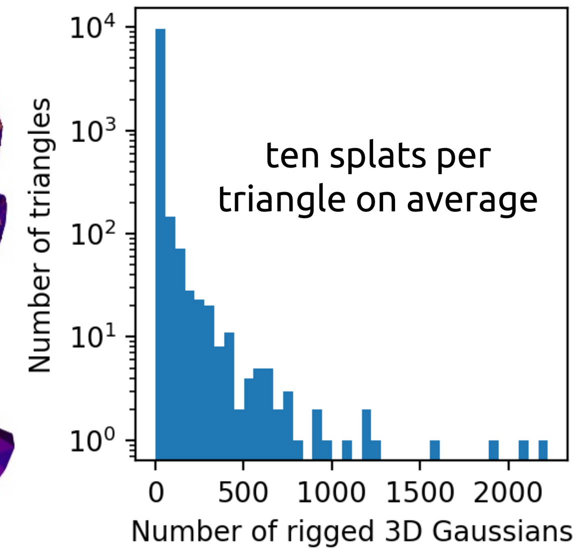
The number of Gaussians :
While the number of Gaussians grows by as much as 1000% during training, the run-time of each optimization iteration increases by less than 100%





Resolution	FPS
401×225	187
802×550	187
1604×1100	156
3208×2200	95
6416×4400	36





	Novel-View			Self-Reenactment		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Ours	28.8	0.883	0.098	25.1	0.853	0.101
w/o ADC	26.8	0.854	0.206	25.1	0.860	0.183
w/o $\mathcal{L}_{\text{scaling}}$	28.0	0.877	0.114	24.9	0.852	0.109
w/o $\epsilon_{\text{scaling}}$	25.0	0.833	0.195	24.1	0.843	0.176
w/o $\mathcal{L}_{\text{position}}$	29.7	0.894	0.091	24.9	0.851	0.096
w/o $\epsilon_{\text{position}}$	28.7	0.882	0.105	25.0	0.855	0.106
w/o FLAME ft.	26.1	0.855	0.131	25.5	0.862	0.124