GaussianVideo: Efficient Video Representation and Compression by Gaussian Splatting

Supplementary Material

A. Appendix

A.1. Experimental Setting

Datasets. We evaluate our method on two video datasets: DAVIS and UVG, as summarized in Tab. I. DAVIS consists of various video sequences with a resolution of 1920×960 and frame counts ranging from 40 to 104. For instance, sequences such as *Blackswan* and *Cows* contain 50 and 104 frames, respectively. UVG, on the other hand, provides longer video sequences with a consistent resolution of 1920×960 and up to 600 frames, as seen in examples like *Beauty* and *ReadySteadyGo*. These datasets offer a diverse range of content for evaluating the scalability and performance of our method.

Dataset	Video	Resolution	Number of Frames
DAVIS	Blackswan	1920 x 960	50
	Bmx-trees	1920 x 960	80
	Boat	1920 x 960	75
	Breakdance	1920 x 960	84
	Camel	1920 x 960	90
	Car-roundabout	1920 x 960	75
	Car-shadow	1920 x 960	40
	Cows	1920 x 960	104
	Dancejump	1920 x 960	60
	Dog	1920 x 960	60
UVG	Beauty	1920 x 960	600
	Bosphorus	1920 x 960	600
	HoneyBee	1920 x 960	600
	Jockey	1920 x 960	600
	ReadySteadyGo	1920 x 960	600
	ShakeNDry	1920 x 960	300
	YachtRide	1920 x 960	600

Table I. Video datasets with resolution and number of frames.

Implementation Details. The details of hyperparameters are summarized in Tab. II. For videos with a resolution of 640×1280, the $0.35M$ model uses 31,024 Gaussians with resolutions of 16×16 for the x, y axes and 8 for t. Two multi-resolution planes are used with scaling ratios of 1 and 2. Enlarging the model size to $0.75M$, the number of Gaussians rises to 50,034, and the resolutions are set to 32×32 for x, y and 16 for t, maintaining the same multi-resolution configuration. For the $1.5M$ model, we use 55,734 Gaussians with resolutions of 32×32 for x, y and 8 for t, extending to three multi-resolution planes with ratios of $1, 2, 4$. Finally, the $3.0M$ model uses 81,954 Gaussians with resolutions of 48×48 for x, y and 12 for t, retaining the same multi-resolution configuration as the $1.5M$ model.

For videos with a resolution of 960×1920, the $0.35M$ model is configured with 38,704 Gaussians and resolutions of 8×8 for x, y and 4 for t, with two multi-resolution planes using scaling ratios of 1 and 2. The $0.75M$ model increases the number of Gaussians to 60,544 and resolutions to 32×32 for x, y and 8 for t. For the $1.5M$ model, we configure it with 68,070 Gaussians and resolutions of 32×32 for x, y and 6 for t, using three multi-resolution planes with ratios of $1, 2, 4$. Lastly, the $3.0M$ model employs 99,666 Gaussians with resolutions of 48×48 for x, y and 10 for t, maintaining the same multi-resolution configuration as the $1.5M$ model.

As the video size increases, the number of Gaussians is scaled accordingly to effectively handle the higher resolution. By increasing the number of Gaussians and their respective resolutions, our method is able to capture the additional spatio-temporal details introduced by larger video dimensions, ensuring robust performance across varying resolutions.

Video size	size (M)	Num of G	x	y	t	ratio
640×1280	0.35	31,024	16	16	8	1,2
640×1280	0.75	50,034	32	32	16	1,2
640×1280	1.5	55,734	32	32	8	1,2,4
640×1280	3	81,954	48	48	12	1,2,4
960×1920	0.35	38,704	8	8	4	1,2
960×1920	0.75	60,544	32	32	8	1,2
960×1920	1.5	68,070	32	32	6	1,2,4
960×1920	3	99,666	48	48	10	1,2,4

Table II. **GaussianVideo** architecture details.

A.2. More Qualitative Results

We present additional qualitative results on a broader set of videos, demonstrating the effectiveness of our method in capturing fine-grained details. Specifically, as shown in Fig. I, in the *Breakdance* video, our model distinctly reconstructs the lettering on the T-shirt, which is not well-represented by other models. Similarly, in the *Car-roundabout* video, unlike other methods, our model accurately reconstructs the "P" sign and the structural shapes. Moreover, in the *Bmx* and *Car-shadow* videos, the details of the wheels are more precisely captured by our method. Notably, in *Car-shadow*, the shadows are faithfully reproduced, showcasing our model's ability to handle subtle visual features. These qualitative results highlight the capability of Gaussian representations to better capture fine-grained details compared to other approaches.

Figure I. Qualitative comparison of different models on videos from the DAVIS dataset, including *Breakdance*, *Camel*, *Bmx*, *Car-roundabout*, *Car-shadow*, and *Dancejump*.