Momentum-GS: Momentum Gaussian Self-Distillation for High-Quality Large Scene Reconstruction

Supplementary Material

1. More Details

Scene partition. (1) Criteria: The scene is first equally divided along the x-axis and then along the z-axis, with each block having the same area. Corresponding views are selected based on visibility. (2) Initialization: Each block is initialized from the same point cloud generated by COLMAP, but only the assigned part and overlapping boundary are reconstructed. (3) Views selection: Views at the boundaries are selected based on visibility, and each block reconstructs an extended region to ensure better reconstruction quality at the boundary area.

Motivation of momentum updates. The momentum-based update provides stable, global guidance, allowing each block's Gaussian decoder to effectively leverage the broader scene context, thereby significantly enhancing reconstruction consistency. As demonstrated in Table 3, using a momentum value of 0.9 outperforms a setting without momentum updates.

2. More Ablation Study

Effectiveness of self-distillation. As shown in Table 1, we performed additional experiments to validate the effectiveness of our self-distillation approach: (1) As shown in setting (b), extending parallel training to 8 blocks with 8 GPUs improved the reconstruction quality. (2) Alternating training across blocks every 500 iterations, using 4 GPUs to train 8 blocks in parallel (setting (c)), slightly decreased the reconstruction quality compared with setting (b). (3) Incorporating our momentum-based self-distillation into setting (c) enhanced the reconstruction quality (setting (d)), clearly demonstrating the effectiveness of our proposed method.

Training strategy	#Block	#GPU	PSNR ↑	SSIM ↑	LPIPS ↓
(a) w/ Parallel training	4	4	23.10	0.790	0.221
(b) w/ Parallel training	8	8	23.34	0.800	0.210
(c) w/ Parallel training (alternating)	8	4	23.17	0.797	0.211
(d) w/ momentum self-distill.	8	4	23.56	0.806	0.205
(e) Full	8	4	**23.65**	**0.813**	**0.194**

Table 1. Ablation study on different training strategies.

The weight of consistency loss. An ablation study is performed to evaluate the impact of the consistency loss weight $\lambda_{consistency}$. As reported in Table 2, the results indicate that model performance remains stable across a wide range of $\lambda_{consistency}$ values.

Momentum value. We ablated momentum value m and Table 3 shows that our model is robust to variations. The reconstruction quality show minimal differences, with the best performance achieved at m=0.9 (our default setting).

Scene	*Building*			*Rubble*		
$\lambda_{consistency}$	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
1	23.53	0.808	0.201	26.51	0.816	0.210
10	23.63	0.810	0.200	26.62	0.821	0.204
50	**23.65**	**0.813**	**0.194**	26.66	0.826	0.200
100	23.63	0.810	0.197	**26.69**	**0.829**	**0.198**

Table 2. Ablation study on $\lambda_{consistency}$.

Momentum values	PSNR ↑	SSIM ↑	LPIPS ↓
0.0	23.44	0.806	0.203
0.5	23.59	0.808	0.201
0.7	23.62	0.810	0.198
0.9 (default)	23.65	0.813	0.194
0.95	23.50	0.806	0.201
0.99	22.06	0.741	0.254

Table 3. Comparison between different momentum values.

3. Quantitative Evaluation

VRAM. We report the peak VRAM usage during inference across five large-scale scenes, as shown in Table 4. Despite achieving superior reconstruction quality, our method requires less VRAM compared to the purely 3DGS-based approach. The VRAM usage, measured in MB, highlights the efficiency of our method. Notably, as scene complexity increases (*e.g.*, in MatrixCity), the advantages of our method become even more pronounced.

Scene	Building	Rubble	Residence	Sci-Art	MatrixCity
CityGaussian	8977	5527	6494	**2726**	14677
Momentum-GS (Ours)	**5830**	**4106**	**6419**	6647	**4616**

Table 4. Peak VRAM usage (in MB) during inference.

Storage. We report the storage usage across five large-scale scenes, as shown in Table 5. Leveraging our hybrid representation, our method significantly reduces the number of parameters required for storage compared to purely 3DGS-based methods. This reduction is especially notable in larger and more complex scenes, such as MatrixCity, where the storage savings are most substantial. Notably, as scene complexity increases (*e.g.*, in MatrixCity), the advantages

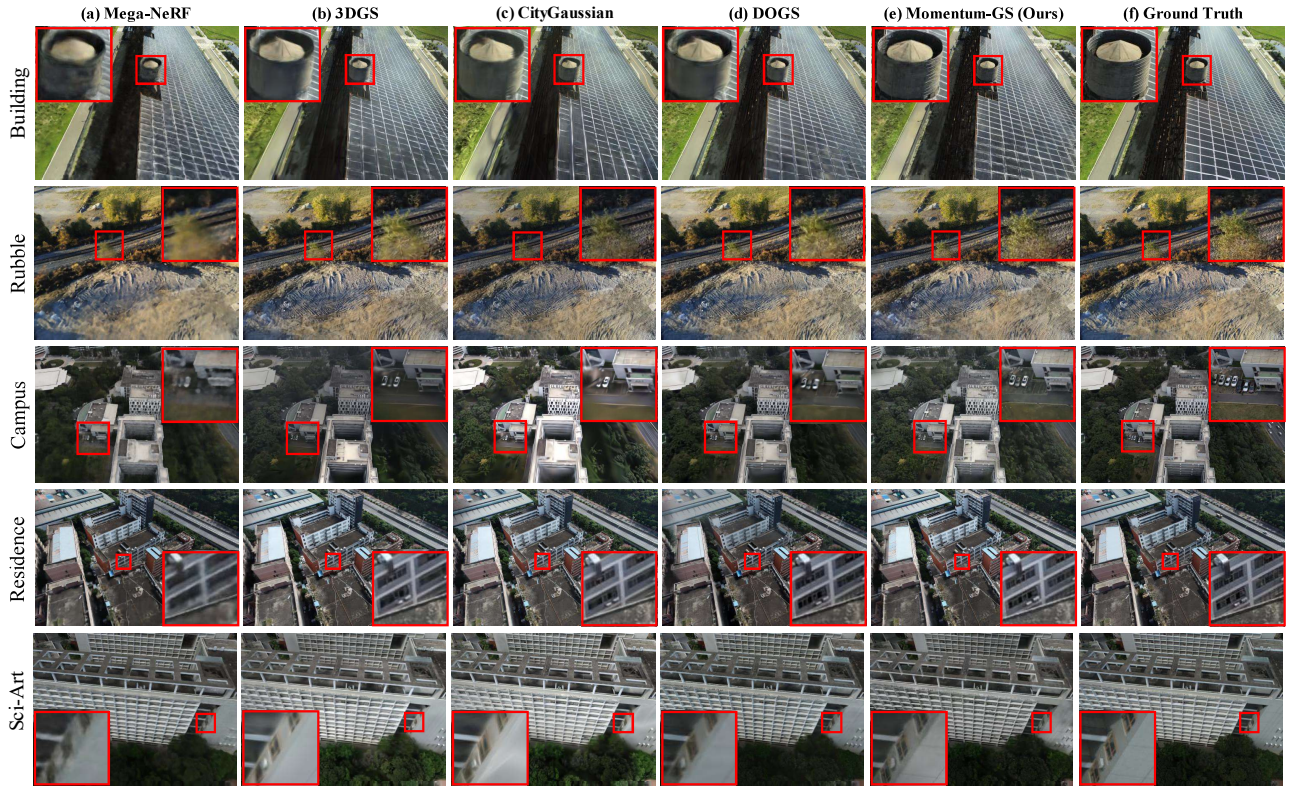

Figure 1. Qualitative comparisons of our Momentum-GS and prior methods across four large-scale scenes.

of our method become even more pronounced, demonstrating its effectiveness in handling challenging scenarios. For clarity and consistency, storage usage is reported in GB.

Scene	Building	Rubble	Residence	Sci-Art	MatrixCity
CityGaussian	3.07	2.22	2.49	**0.88**	5.40
Momentum-GS (Ours)	**2.45** (20.2%↓)	**1.50** (32.7%↓)	**2.00** (19.7%↓)	0.97	**2.08** (61.5%↓)

Table 5. Storage usage (in GB).

Number of primitives. We report the number of primitives across five large-scale scenes, as shown in Table 6.

Scene	Building	Rubble	Residence	Sci-Art	MatrixCity
Primitives	8.33M	5.09M	6.79M	3.30M	7.08M

Table 6. Primitives counts for each scene.

Comparison of different implementations of VastGaussian. We further compare our method with the unofficial implementation of VastGaussian in Table 7, which demonstrates improved performance over the results reported in DOGS.

Scene	*Building*			*Rubble*		
Metrics	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
VastGaussian (DOGS version)	21.80	0.728	0.225	25.20	0.742	0.264
VastGaussian (Unofficial)	22.49	0.742	0.208	25.64	0.760	0.202
Momentum-GS (Ours)	**23.65**	**0.813**	**0.194**	**26.66**	**0.826**	**0.200**

Table 7. Comparison of different implementations of VastGaussian.

4. More Visual Comparisons

We provide additional visual comparisons for the Building, Rubble, Residence, and Sci-Art scenes in Figure 1. Our method consistently reconstructs finer details across these scenes. Notably, our approach demonstrates a superior ability to reconstruct luminance, as illustrated by the Sci-Art example shown in Figure 1. While NeRF-based methods are capable of capturing luminance by leveraging neural networks to learn global features such as lighting, they tend to produce blurrier results compared to 3DGS-based methods. This underscores the effectiveness of our hybrid representation, which combines the strengths of both NeRF-based and 3DGS-based approaches.