Energy-GS: Supplementary Material

Yu Gao Lutong Su Ruixiang Huang Tianji Jiang Jiadong Tang Yufeng Yue Yi Yang*
Beijing Institute of Technology

1. Experiments

1.1. Performance on Benchmark Datasets

The remaining experimental results on the synthetic and Mip-NeRF 360 datasets are presented in Figure 1, Figure 2 and Table 1, Table 2. As shown, our method achieves rendering and pose estimation performance consistent with the conclusions reported in the main paper across all additional scenes.

We also provide the complete joint optimization training process to help readers gain a more intuitive understanding of our method. As shown in **Video 1**, the left part of the video presents the results of the original 3DGS with only pose gradients added for joint optimization, while the right part shows our Energy-GS. From the video, it can be seen that the ground-truth images used by our method are not the original RGB images, but progressively modified images controlled by energy. Compared to the original method, our coarse-to-fine strategy ensures rapid convergence of camera poses.

Another difference from the original 3DGS is that, during the initial stage of pose alignment, the positions of Gaussian primitives are not optimized. The fixed number of Gaussian primitives guarantees the stability of gradient optimization. The optimized point clouds corresponding to the Gaussian primitives for both methods are displayed at the bottom-right of each respective method. These two strategies correspond to the two main contributions proposed in our paper.

Additionally, To ensure that the estimated camera poses are comparable with the ground-truth poses, it is necessary to align their global coordinate system, otherwise, the evaluation metrics would be ambiguous. We fix the first camera of each scene and keep its parameters non-learnable during joint optimization. This ensures that all optimized poses are anchored to the ground-truth coordinate frame.

1.2. Signal Alignment Task for 1DGS

The signal alignment task can be viewed as a simplified version of projecting 3DGS onto a 1D space, as formulated in Eq.2 of the main paper. The complete signal alignment process and visualization results are provided in Supplementary **Video 2**. Compared with the results shown in Figure. 4 of the main paper, the submitted video includes more details, specifically including:

- **Training Process:** The full optimization trajectory of the cropped signal during training.
- **Rendering Total:** The full optimization trajectory of the complete signal during training.
- **Alignment Rect:** The entire alignment procedure applied to the cropped signal during training.
- **GS Distribution:** The evolution of the exponent term of each Gaussian primitive during training.
- **Energy Signal:** The ground-truth signal after applying the energy-based control strategy. If the strategy is disabled, this signal remains identical to the original cropped input.
- **GS OPA:** The opacity of each Gaussian primitive throughout training.
- **GS VAL:** The value of each Gaussian primitive during training, corresponding to its color representation in 3DGS (like the spherical harmonics coefficients).

The video presents ablation studies under different combinations of strategies. From the complete training trajectories, it is evident that both proposed strategies are effective, and using only one of them is insufficient to achieve the desired performance. As shown in the "Rendering Total" subfigure, high-quality reconstruction of the full signal is possible only when the cropped signal is properly aligned. In the context of 3D space, this implies that novel-view rendering quality is accurate only when camera poses are aligned.

In addition, across all four experiments, the estimated cropped signals (sub-figure "Training Process") appear accurate when considered in isolation. However, high-quality local signal rendering does not necessarily imply that the signal positions are aligned, as observed in Exp. (a): Gaussian Splatting. This further demonstrates that relying solely on signal values as constraints, which corresponds to using only RGB images as supervision in 3DGS, is a highly challenging task in joint optimization. Our method therefore provides valuable insight into addressing this problem.

1.3. Camera Pose Gradients

There are three backpropagation pathways for the camera pose gradients, ordered from strongest to weakest influence

as follows:

- **First:**

$$Loss \rightarrow \alpha_i \rightarrow T \rightarrow se3 \tag{1}$$

- **Second:**

$$Loss \rightarrow acc_i \rightarrow T \rightarrow se3 \tag{2}$$

- **Third:**

$$Loss \rightarrow c_i \rightarrow T \rightarrow se3 \tag{3}$$

α_i is given by evaluating a 2D Gaussian with covariance Σ' multiplied with a learned per-point opacity. acc_i represents the accumulated transmittance, c_i is the color of the 3D Gaussian, T denotes the camera pose, and $se3$ refers to the learnable pose parameters in the Lie algebra space.

First Pathway. The detailed form of the first path consists of two components:

$$\begin{cases} Loss \rightarrow \alpha_i \rightarrow \mu \rightarrow T \rightarrow se3 \\ Loss \rightarrow \alpha_i \rightarrow \Sigma' \rightarrow T \rightarrow se3 \end{cases} \tag{4}$$

where μ and Σ' denote the mean (position) and covariance of the projected 2D Gaussian, respectively.

The gradient of $Loss \rightarrow \alpha_i$ is:

$$\frac{\partial L}{\partial \alpha_i} = \frac{\partial L}{\partial C}\frac{\partial C}{\partial \alpha_i} \tag{5}$$

$$\frac{\partial L}{\partial \alpha_i} = \frac{\partial L}{\partial C}(c_i \cdot acc_i + \sum_{k>i} c_k \cdot \alpha_k \cdot \frac{\partial acc_k}{\partial \alpha_i}) \tag{6}$$

$$\frac{\partial L}{\partial \alpha_i} = \frac{\partial L}{\partial C}(c_i \cdot acc_i - \sum_{k>i} \cdot \frac{acc_k \cdot c_k \cdot \alpha_k}{1-\alpha_i}) \tag{7}$$

where:

$$acc_k = \prod_{j<k}(1-\alpha_j) \tag{8}$$

The gradient of $\alpha_i \rightarrow \mu$ is:

$$\frac{\partial \alpha_i}{\partial \mu} = \frac{\partial \alpha_i}{\partial G_i}\frac{\partial G_i}{\partial \mu} \tag{9}$$

$$\frac{\partial \alpha_i}{\partial \mu} = opa_i \cdot \frac{\partial G_i}{\partial \mu} \tag{10}$$

$$\frac{\partial \alpha_i}{\partial \mu} = opa_i \cdot G_i \cdot \Sigma'^{-1} \cdot (p-\mu) \tag{11}$$

where:

$$G_i = e^{(-\frac{1}{2}\cdot r_i^T \cdot \Sigma'^{-1} \cdot r_i)} \tag{12}$$

$$r = p - \mu \tag{13}$$

p is the pixel coordinate and opa_i is the learnable opacity of the Gaussian.

The gradient of $\alpha_i \rightarrow \Sigma'$ is:

$$\frac{\partial \alpha_i}{\partial \Sigma'} = \frac{\partial \alpha_i}{\partial G_i}\frac{\partial G_i}{\partial \Sigma'} \tag{14}$$

$$\frac{\partial \alpha_i}{\partial \Sigma'} = opa_i \cdot \frac{\partial G_i}{\partial \Sigma'} \tag{15}$$

$$\frac{\partial \alpha_i}{\partial \Sigma'} = -\frac{1}{2} \cdot opa_i \cdot G_i \cdot \Sigma'^{-1} \cdot r_i \cdot r_i^T \cdot \Sigma'^{-1} \tag{16}$$

Next, we consider the gradient mapping from 2D to 3D, beginning with the derivation of $\mu \rightarrow T \rightarrow se3$.

We have the 3D to 2D projection equation:

$$\mu = \pi(T \cdot \mu_g) \tag{17}$$

where π is the perspective projection, μ_g denotes the 3D Gaussian mean (position). We have:

$$\frac{\partial \mu}{\partial se3} = \frac{\partial \mu}{\partial T}\frac{\partial T}{\partial se3} = \frac{\partial \mu}{\partial (T \cdot \mu_g)}\frac{\partial (T \cdot \mu_g)}{\partial T}\frac{\partial T}{\partial se3} \tag{18}$$

The Jacobian matrix of the perspective projection is given by:

$$\frac{\partial \mu}{\partial (T \cdot \mu_g)} = J_\pi = \begin{bmatrix} 1/Z & 0 & -X/Z^2 \\ 0 & 1/Z & -Y/Z^2 \end{bmatrix} \tag{19}$$

Where:

$$\begin{bmatrix} X & Y & Z \end{bmatrix}^T = T \cdot \mu_g \tag{20}$$

The Jacobian matrix of the world to camera transformation with respect to the Lie algebra $se3$ is:

$$J_T = \begin{bmatrix} I & -[T \cdot \mu_g]_\wedge \end{bmatrix} \tag{21}$$

Here, I denotes the 3×3 identity matrix, and $[\cdot]_\wedge$ represents the skew-symmetric matrix of a vector. We have:

$$\frac{\partial \mu}{\partial se3} = J_\pi \cdot J_T \tag{22}$$

Next, we derive the gradient for $2D \rightarrow 3D$ mapping from $\Sigma' \rightarrow T \rightarrow se3$. According to the 2D covariance provided in EWA Splatting:

$$\Sigma' = JW\Sigma W^T J^T \tag{23}$$

Σ is the 3D covariance of the Gaussian primitive in the world coordinate system, W is the rotation part of the world to camera transformation T, and J is the Jacobian matrix of the perspective projection. We have:

$$\frac{\partial \Sigma'}{\partial se3} = \frac{\partial (JW\Sigma W^T J^T)}{\partial W} \frac{\partial W}{\partial se3} \tag{24}$$

The derivation of the first part requires matrix differential calculus. Let $A = W\Sigma W^T$, then:

$$\Sigma' = JAJ^T \tag{25}$$

$$d\Sigma' = J(dA)J^T \tag{26}$$

$$dA = (dW)\Sigma W^T + W\Sigma(dW^T) \tag{27}$$

Since Σ is a symmetric matrix and $dW^T = (dW)^T$, the above formula becomes:

$$dA = (dW)\Sigma W^T + W\Sigma(dW)^T \tag{28}$$

Substituting the above expression into Eq 26, we have:

$$d\Sigma' = J(dW)\Sigma W^T J^T + JW\Sigma(dW)^T J^T \tag{29}$$

According to the rules of matrix calculus, we have:

$$\frac{\partial \Sigma'}{\partial W} = J \otimes (JW\Sigma) + (JW\Sigma) \otimes J \tag{30}$$

Where $\otimes$ denotes the Kronecker product. For the part of $\partial W / \partial \mathrm{se3}$, the translational part in $se3$ does not participate. Only the rotational parameters ξ in $se3$ are considered, as only the rotation affects the gradient:

$$\frac{\partial W}{\partial \xi} = I - \frac{1-\cos\theta}{\theta^2}[\xi]_\times + \frac{\theta - \sin\theta}{\theta^3}[\xi]_\times^2 \tag{31}$$

Finally, we combine all the above results, for the chain $Loss \rightarrow \alpha_i \rightarrow T \rightarrow se3$:

$$\frac{\partial L_1}{\partial se3} = \frac{\partial L}{\partial C}\frac{\partial C}{\partial \alpha_i}\left(\frac{\partial \alpha_i}{\partial \Sigma'}\frac{\partial \Sigma'}{\partial W}\frac{\partial W}{\partial se3} + \frac{\partial \alpha_i}{\partial G}\frac{\partial G}{\partial \mu}\frac{\partial \mu}{\partial se3}\right) \tag{32}$$

Second Pathway. The second pathway is $Loss \rightarrow acc_i \rightarrow T \rightarrow se3$. For the part $Loss \rightarrow acc_i$ is:

$$\frac{\partial L}{\partial acc_i} = \frac{\partial L}{\partial C}\frac{\partial C}{\partial acc_i} \tag{33}$$

$$\frac{\partial L}{\partial acc_i} = \frac{\partial L}{\partial C} \cdot c_i \cdot \alpha_i \tag{34}$$

The derivative of acc_i with respect to all previous α_j $(j < i)$ is:

$$\frac{\partial acc_i}{\partial \alpha_j} = \frac{-acc_i}{1-\alpha_j} \tag{35}$$

The effect on $se3$ is:

$$\frac{\partial \alpha_j}{\partial se3} = \frac{\partial \alpha_j}{\partial T}\frac{\partial T}{\partial se3} \tag{36}$$

$$\frac{\partial \alpha_j}{\partial T}\frac{\partial T}{\partial se3} = \frac{\partial \alpha_j}{\partial \Sigma'}\frac{\partial \Sigma'}{\partial W}\frac{\partial W}{\partial se3} + \frac{\partial \alpha_j}{\partial G}\frac{\partial G}{\partial \mu}\frac{\partial \mu}{\partial se3} \tag{37}$$

Eq 37 has already been derived in the first pathway. Finally, by consolidating all the above results, for the pathway $Loss \rightarrow acc_i \rightarrow T \rightarrow se3$, we have

$$\frac{\partial L_2}{\partial se3} = \frac{\partial L}{\partial C}\frac{\partial C}{\partial acc_i}\frac{\partial acc_i}{\partial \alpha_j}\frac{\partial \alpha_j}{\partial T}\frac{\partial T}{\partial se3} \tag{38}$$

Third Pathway. In fact, this gradient path is not utilized in our work. We observed that the spherical harmonics exhibit relatively smooth changes with respect to view directions, and minor color variations between adjacent viewpoints are almost immediately compensated by the learnable variables of the spherical harmonics. As a result, when the camera poses converge near the ground truth, this path provides almost no effective gradient. Moreover, when the color learning rate is relatively high, it can even induce frequent pose fluctuations around the ground-truth poses.

Naturally, ignoring this path introduces new challenges. Specifically, when reconstructing surfaces with specular reflections or regions with highly sensitive color variations, our method may struggle to achieve optimal results, as seen in the materials scene of the synthetic dataset.

1.4. Limitations

While Energy-GS effectively enhances joint optimization performance, our method has several limitations. The first is its sensitivity to the number of initial Gaussian primitives. During the pose-alignment stage (before densification begins, as described in Sec. 5 of the main paper), both the positions and the number of Gaussians remain fixed. Consequently, it is critical that these initial Gaussians can coarsely represent the scene. If too few Gaussians are initialized, their representational capacity is insufficient, leading to rendering collapse and erroneous pose drift. Conversely, an excessive number of Gaussians can lead to overfitting. Free Gaussians outside the shared field of view may encode extraneous scene details, resulting in under-optimization of the Gaussian set. As a result, camera pose estimation may stagnate prematurely and converge to a local optimum.

The second limitation is its sensitivity to hyperparameters, particularly the trade-off between the image-energy scheduling rate and the pose learning rate. As shown in Eq. 9 of the main paper, the energy-weighting term is adjusted over training steps according to a predefined schedule, but no feedback mechanism is provided to adapt it to the state of scene reconstruction or pose optimization. In practice, the convergence behavior of pose optimization varies with

Table 1. Quantitative rendering results of different methods on the remaining scenes in the datasets.

Scene	BARF			SC-NeRF			CF-GS			3R-GS			3DGS			Ours		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
drums	22.71	0.876	0.179	10.45	0.674	0.590	0.739	0.001	0.927	14.75	0.839	0.332	11.40	0.684	0.531	22.75	0.854	0.161
materials	26.40	0.907	0.136	8.42	0.615	0.596	4.13	0.377	0.829	12.44	0.757	0.362	8.72	0.750	0.450	25.15	0.898	0.131
mic	28.53	0.952	0.095	11.72	0.754	0.516	4.90	0.456	0.680	14.87	0.870	0.296	13.04	0.884	0.298	28.61	0.940	0.036
ship	25.46	0.806	0.274	8.72	0.563	0.626	-	-	-	12.28	0.675	0.561	6.71	0.541	0.678	23.51	0.742	0.311
bonsai	12.26	0.286	0.890	13.92	0.321	0.799	9.36	0.022	0.973	9.44	0.051	0.841	10.64	0.307	0.833	18.06	0.487	0.186
flowers	10.50	0.162	0.997	12.70	0.176	0.896	7.53	0.022	1.060	21.26	0.584	0.261	14.09	0.171	0.764	19.78	0.367	0.351
kitchen	10.40	0.203	0.913	13.70	0.303	0.858	5.80	0.011	1.040	13.55	0.372	0.965	12.77	0.340	0.730	17.35	0.385	0.285
room	11.59	0.321	0.752	12.84	0.350	0.727	8.63	0.046	0.953	10.25	0.297	0.892	12.78	0.385	0.633	21.17	0.651	0.197
treehill	12.24	0.244	0.956	14.02	0.292	0.892	8.44	0.030	1.029	22.86	0.664	0.254	17.36	0.373	0.590	18.83	0.301	0.334

Table 2. Quantitative pose estimation results of different methods on the remaining scenes in the datasets.

Scene	BARF		SC-NeRF		CF-GS		3R-GS		3DGS		Ours	
	Rotation(°)↓	ATE(m)↓	Rotation(°)↓	ATE(m)↓	Rotation(°)↓	ATE(m)↓	Rotation(°)↓	ATE(m)↓	Rotation(°)↓	ATE(m)↓	Rotation(°)↓	ATE(m)↓
drums	0.788	0.028	7.897	0.095	92.155	1.155	82.329	0.936	29.661	0.244	0.439	0.017
materials	0.084	0.019	7.903	0.095	112.102	1.111	111.144	1.038	39.414	0.323	0.114	0.031
mic	0.832	0.027	7.954	0.101	101.441	1.414	110.429	1.130	14.197	0.194	0.977	0.030
ship	0.361	0.021	7.876	0.099	-	-	77.974	1.053	40.364	0.275	0.401	0.010
bonsai	10.048	0.131	8.102	0.091	102.818	1.032	21.972	1.380	6.921	0.101	0.745	0.019
flowers	1.230	0.059	7.978	0.095	122.472	1.016	0.151	0.026	2.208	0.080	0.367	0.016
kitchen	0.515	0.079	7.947	0.093	124.418	0.715	6.506	1.128	5.891	0.123	0.781	0.016
room	1.567	0.036	7.888	0.073	115.110	0.840	106.516	0.461	6.185	0.070	0.890	0.028
treehill	2.958	0.077	8.119	0.085	139.615	0.946	0.585	0.012	2.309	0.056	0.789	0.020

different initial poses. Based on our experience, when the energy weight is increased too rapidly while the pose learning rate is relatively small, the camera poses tend to stop improving prematurely and fail to reach the ground truth. Conversely, when the energy schedule is too slow but the pose learning rate is large, the poses may align too early, then begin to oscillate, and eventually drift away from the ground truth without recovering.

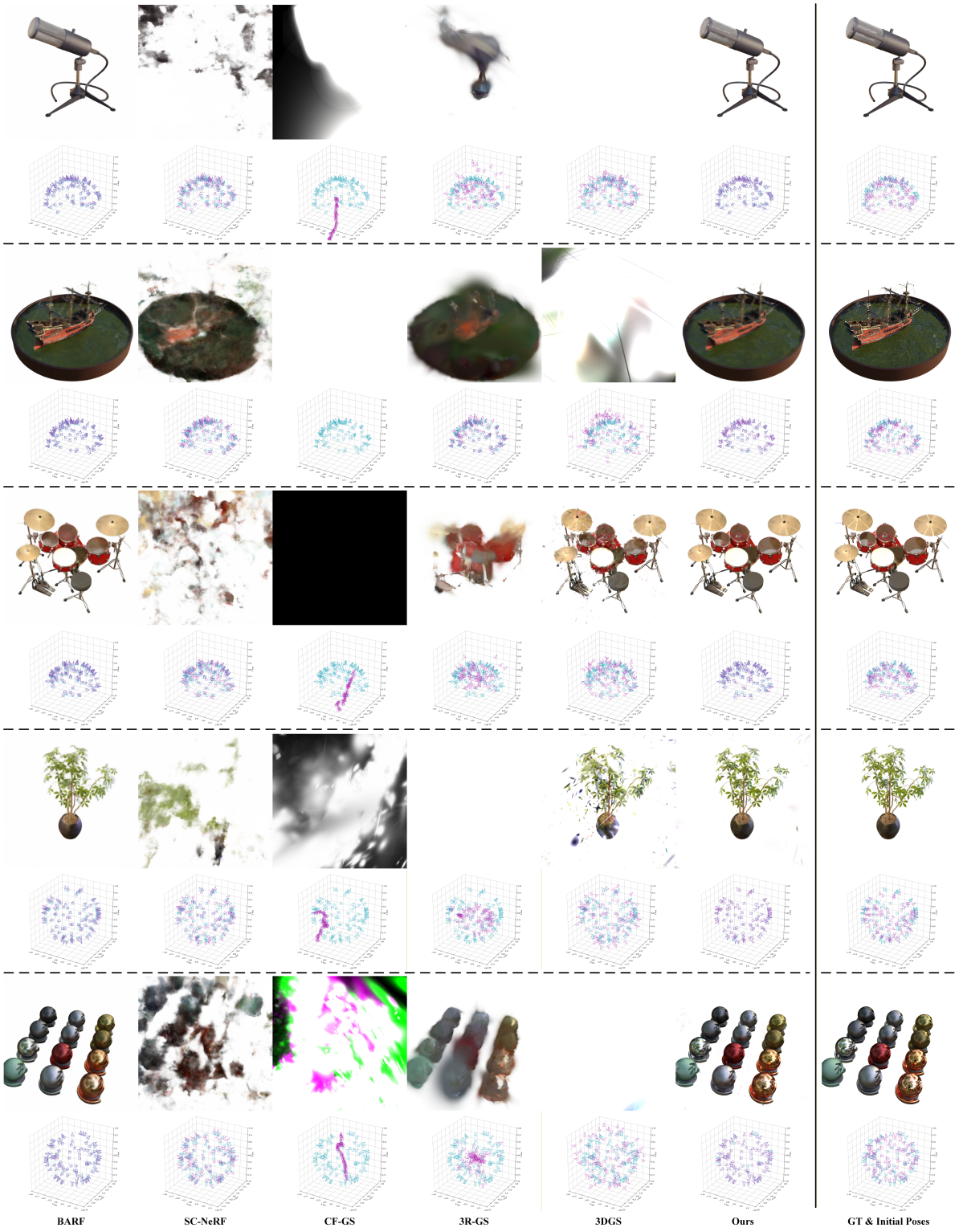

Figure 1. Visual rendering results and camera pose visualization results of different methods on the remaining scenes of the synthetic dataset.

Figure 2. Visual rendering results and camera pose visualization results of different methods on the remaining scenes of the Mip-NeRF 360 dataset.