



Gaussian Herding across Pens: An Optimal Transport Perspective on Global Gaussian Reduction for 3DGS

Tao Wang^{1*}, Mengyu Li^{2*}, Geduo Zeng¹, Cheng Meng^{1†}, Qiong Zhang^{1†}

¹Renmin University of China, ² Tsinghua University

*equal contribution, [†]corresponding

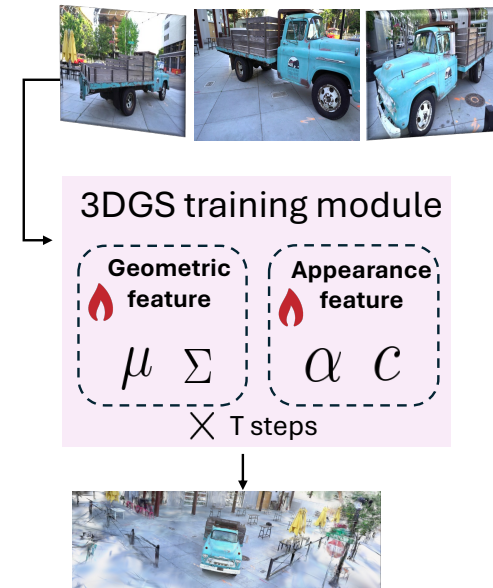
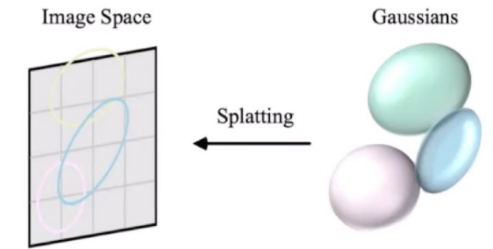


Background

🌐 3DGS represents 3D scenes with many tiny Gaussian primitives, enabling fast, photorealistic rendering of complex environments.

⚠️ However, 3DGS often contains **millions of redundant** Gaussians, which increase memory usage and slow down rendering.

✂️ One way to address this is through compaction—reducing the number of Gaussians while preserving visual quality.



Existing compaction work

- Pruning based methods:

Compacting by removing low-score Gaussian elements often leads to **geometric distortion**.

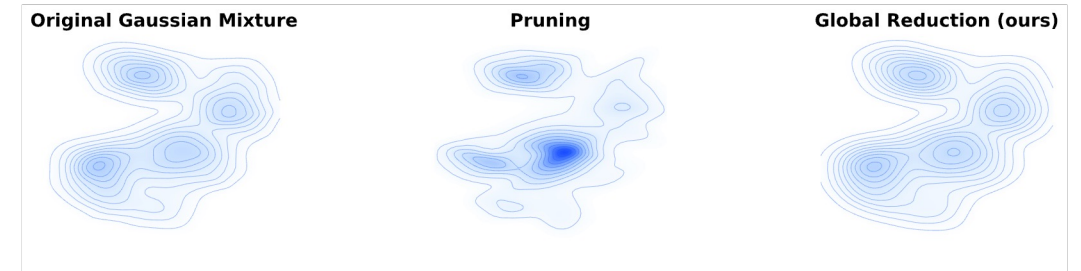
Examples: Light-GS (Fan et al. 2024), PUP-3DGS (Hanson et al. 2024), Tramming-the-fat (Ali et al. 2024)

- End-to-end compaction

Directly training a lighter-weight 3DGS model; however, the **number of Gaussians is typically uncontrollable** and **coupled to the training pipeline**, limiting generalization.

Examples: 3DGS-MCMC (Kheradmand et al. 2024), LocoGS (Shin et al. 2025)

✂ Pruning may break geometric consistency, so we propose a **global** approach that **preserves both shape and appearance**.



Gaussian Mixture Reduction method

- We interpret the 3DGS model as a **Gaussian Mixture Model**:

$$\phi_n(x) = \sum_{i=1}^n \alpha_i \phi(x; \mu_i, \Sigma_i),$$

The compacted 3DGS model is also a **Gaussian mixture**:

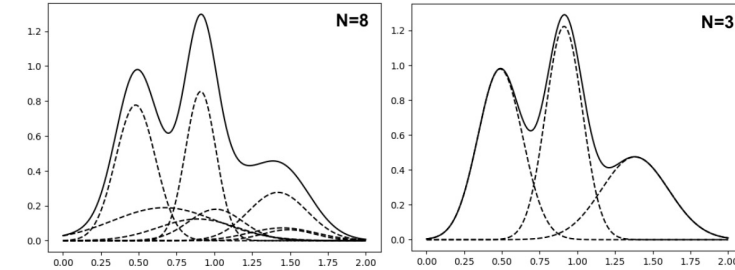
$$\phi'_m(x) = \sum_{j=1}^m \alpha'_j \phi(x; \mu'_j, \Sigma'_j)$$

- Find $\phi'_m(x)$ such that $\mathcal{T}_c(\phi_n, \phi'_m)$ as small as possible

$$\mathcal{T}_c(\phi_n, \phi'_m) = \inf \left\{ \sum_{i=1}^n \sum_{j=1}^m \pi_{ij} c(\phi(\cdot; \mu_i, \Sigma_i), \phi(\cdot; \mu'_j, \Sigma'_j)) : \sum_{j=1}^m \pi_{ij} = \alpha_i, \sum_{i=1}^n \pi_{ij} = \alpha'_j \right\}$$

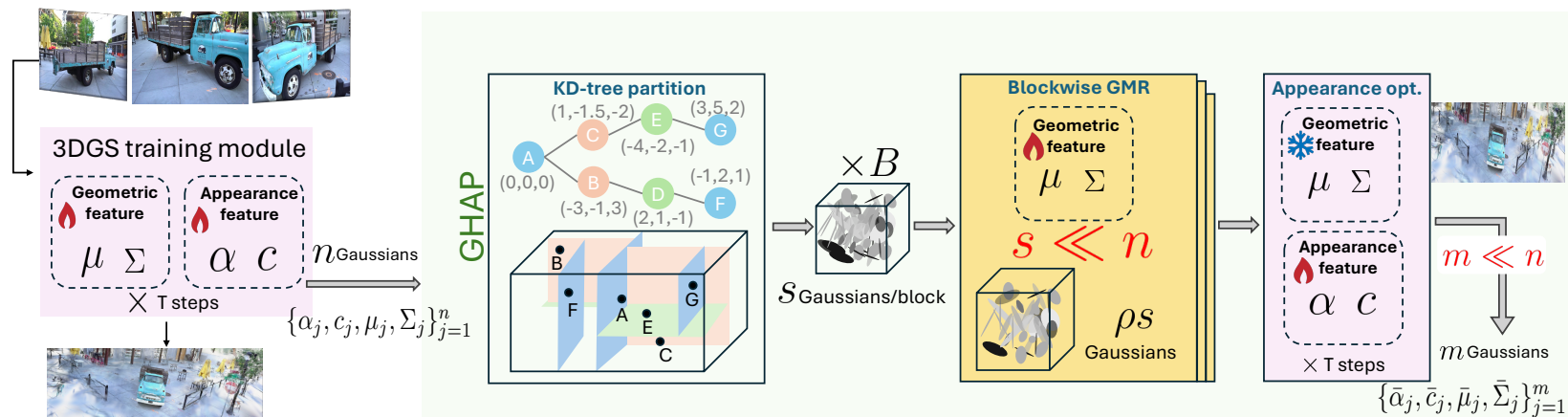
- Why composite transportation divergence?
 - Efficient algorithm. By Zhang et al. the optimization procedure **mimics a k-means**
 - New cost function. Let $c(\phi, \phi') = \|\mu - \bar{\mu}\|_2^2 + \|\Sigma - \bar{\Sigma}\|_F^2$, the one-step update is,

$$\bar{\mu}_j = \frac{\sum_{i \in \mathcal{C}_j} \alpha_i \mu_i}{\sum_{i \in \mathcal{C}_j} \alpha_i}, \quad \bar{\Sigma}_j = \frac{\sum_{i \in \mathcal{C}_j} \alpha_i \Sigma_i}{\sum_{i \in \mathcal{C}_j} \alpha_i}.$$



Example of GMR. The right one has same overall shape as left, but with fewer Gaussians

GHAP Pipeline



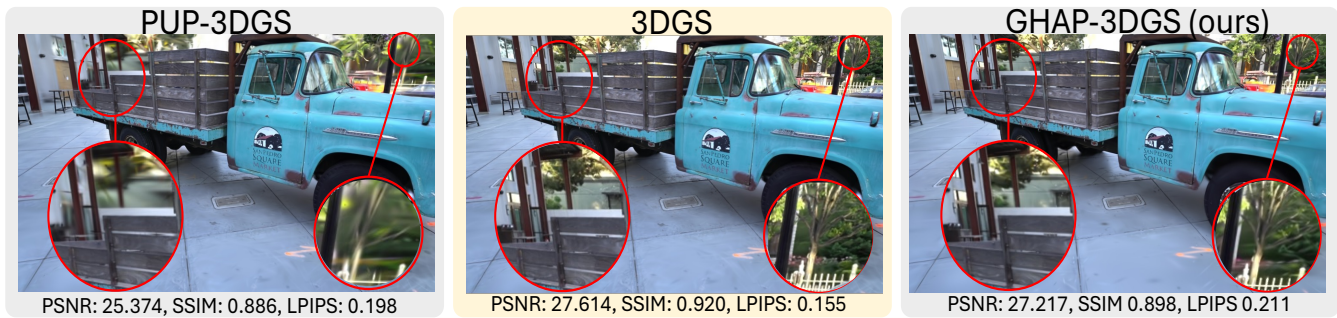
1. GMR within blocks partitioned by KD-tree.

As global GMR is 1) sensitive to outliers 2) computationally expensive 3) lose fine-grained geometric details, we turn to use blockwise GMR

2. Appearance Optimization.

After running GMR, fine-tune all features except the Gaussian shape parameters to adapt to the updated geometry and enhance visual quality

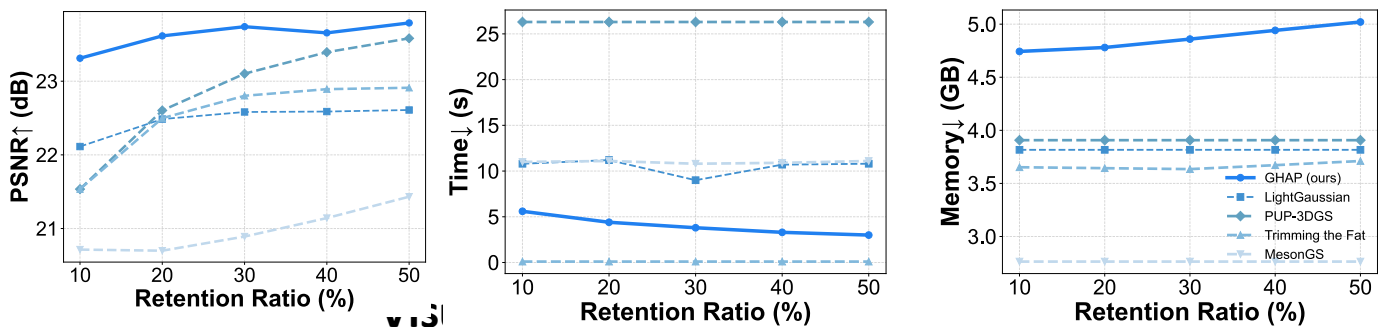
Comparison



GHAP preserves more geometric details.

Method	Tanks&Temples				MipNeRF-360				Deep Blending			
	SSIM↑	PSNR↑	LPIPS↓	k Gaussians	SSIM↑	PSNR↑	LPIPS↓	k Gaussians	SSIM↑	PSNR↑	LPIPS↓	k Gaussians
Vanilla 3DGS	0.853	23.785	0.169	1577	0.813	27.554	0.221	2627	0.907	29.816	0.238	2475
LocoGS	0.843	23.655	0.191	571	0.798	27.049	0.257	674	0.903	29.972	0.261	529
3DGS+GHAP (ours)	0.818	23.312	0.242	157	0.764	26.404	0.314	263	0.905	29.647	0.264	248
LightGaussian	0.756	22.113	0.306	158	0.735	25.674	0.331	263	0.869	28.010	0.327	248
PUP-3DGS	0.767	21.519	0.280	158	0.753	25.332	0.309	262	0.895	29.153	0.274	248
Trimming the Fat	0.776	21.535	0.293	156	0.731	25.255	0.343	263	0.887	28.056	0.302	247
MesonGS	0.811	20.714	0.208	157	0.773	24.924	0.264	263	0.896	28.693	0.264	248
3DGS-MCMC	0.779	22.141	0.282	157	0.763	25.957	0.309	263	0.885	28.976	0.298	248
MiniSplatting	0.799	22.661	0.265	78	0.759	26.022	0.318	111	0.895	29.395	0.289	125
MiniSplatting+GHAP (ours)	0.835	23.232	0.198	79	0.802	27.090	0.250	112	0.909	30.042	0.254	127

GHAP consistently matches or surpasses pruning-based and end-to-end baselines, at a low retention rate (10%)









GHAP outperforms pruning-based approaches across different retention levels, with faster compaction and a slight increase in memory usage.

Conclusion

- We are the first to reinterpret Gaussian primitives in 3DGS as components of a Gaussian mixture and reformulated 3DGS compaction as Gaussian Mixture Reduction
- We also develop a block-wise GMR algorithm guided by a KD-tree, enabling **efficient large-scale scene compaction**.
- Our method is **post-hoc** and **compatible** with any existing 3DGS pipeline

Reference

-  Fan, Zhiwen, et al. "Lightgaussian: Unbounded 3d gaussian compression with 15x reduction and 200+ fps." *Advances in neural information processing systems* 37 (2024): 140138-140158.
-  Hanson, Alex, et al. "Pup 3d-gs: Principled uncertainty pruning for 3d gaussian splatting." *Proceedings of the Computer Vision and Pattern Recognition Conference*. (2025).
-  Ali, Muhammad Salman, et al. "Trimming the fat: Efficient compression of 3d gaussian splats through pruning." *The British Machine Vision Conference* (2024).
-  Kheradmand, Shakiba, et al. "3d gaussian splatting as markov chain monte carlo." *Advances in Neural Information Processing Systems* 37 (2024): 80965-80986.
-  Shin, Seungjoo, Jaesik Park, and Sunghyun Cho. "Locality-aware gaussian compression for fast and high-quality rendering." *International Conference on Learning Representations* (2025).
-  Q. Zhang, A. G. Zhang and J. Chen, "Gaussian Mixture Reduction With Composite Transportation Divergence," in *IEEE Transactions on Information Theory*, vol. 70, no. 7, pp. 5191-5212, July 2024, doi: 10.1109/TIT.2023.3323346