Supplementary Materials for “GAP: Gaussianize Any Point Clouds with Text Guidance”

Weiqi Zhang[1*], Junsheng Zhou[1*†], Haotian Geng[1*], Wenyuan Zhang[1], Yu-Shen Liu[1†]
School of Software, Tsinghua University, Beijing, China[1]
{zwq23, zhou-js24, genght24, zhangwen21}@mails.tsinghua.edu.cn
liuyushen@tsinghua.edu.cn

1. More Implementation Details

Our text-to-image generation pipeline is based on the Stable Diffusion v1.5 model [4] with ControlNet [7] for depth conditioning. Specifically, we utilize the Depth-to-Image model architecture that incorporates depth maps as geometric guidance during the diffusion process. For Gaussian optimization, each viewpoint undergoes 1000 iterations of optimization. During optimization, we set the distance loss weight α to 1.5 and the scale loss weight β to 5. For the Scale Loss, we set the maximum scale threshold τ to $1e-6$. In the spatial-aware Gaussian inpainting stage, we set the number of nearest neighbors $L = 90$ for color diffusion and the radius $\rho = 0.1$ for opacity control. The base opacity value o_0 is set to 5 and the density threshold P_0 is set to 100. All experiments are conducted on NVIDIA RTX3090. The complete pipeline for processing an object takes about 25 minutes.

Adaption of Baselines in Point-to-Gaussian Generation. We extended DreamGaussian, originally designed for image/text-to-3D generation, to support point cloud inputs by replacing its random Gaussian initialization with point cloud-guided initialization. We adapt TriplaneGaussian as our baseline by modifying its original image-conditioned 3DGS generation pipeline. Specifically, we bypass its point cloud decoder for direct point-to-Gaussian conversion and incorporate Stable Diffusion to enable text-to-image generation. DiffGS uses a Gaussian VAE to convert point clouds into Gaussians by querying features from triplanes. However, its lack of text-guided appearance control limited us to conducting only unconditional point-to-gaussian generation experiments.

2. Evaluation Metrics

We employ three complementary metrics to comprehensively evaluate our method’s performance: Fréchet Inception Distance (FID) [6], Kernel Inception Distance (KID) [2], and CLIP Score [3]. For each metric, we render the generated results from fixed viewpoints at 1024×1024 resolution.

2.1. Fréchet Inception Distance (FID)

FID measures the similarity between the distribution of generated images and real images. We compute feature representations using the InceptionV3 [5] network pretrained on ImageNet. The FID score is calculated as:

$$\text{FID} = \|\mu_r - \mu_g\|^2 + \text{Tr}\left(\Sigma_r + \Sigma_g - 2\left(\Sigma_r\Sigma_g\right)^{1/2}\right), \quad (1)$$

where μ_r, Σ_r and μ_g, Σ_g are the mean and covariance matrices of the real and generated feature distributions respectively. Lower FID scores indicate better generation quality.

2.2. Kernel Inception Distance (KID)

KID provides an unbiased estimate of the Maximum Mean Discrepancy (MMD) between real and generated image features. We report KID scores multiplied by 10^3 for better readability. The KID metric is computed as:

$$\begin{aligned} \text{KID} &= \text{MMD}^2(X_r, X_g) \\ &= \frac{1}{n(n-1)}\sum_{i \neq j} k(x_i, x_j) + \frac{1}{m(m-1)}\sum_{i \neq j} k(y_i, y_j) \\ &\quad - \frac{2}{mn}\sum_{i,j} k(x_i, y_j), \end{aligned} \quad (2)$$

where X_r and X_g are real and generated feature sets respectively, and $k(,)$ is the polynomial kernel. Like FID, lower KID scores indicate better quality.

*Equal contribution. † Corresponding authors. This work was supported by National Key R&D Program of China (2022YFC3800600), and the National Natural Science Foundation of China (62272263), and in part by Tsinghua-Kuaishou Institute of Future Media Data.

2.3. CLIP Score

CLIP Score evaluates the semantic alignment between the images rendered from Gaussians and the input text prompts. We use the CLIP ViT-L/14 [3] model to compute the cosine similarity between text and image embeddings. For each generated result, we average the CLIP scores across all rendered views. Higher CLIP scores indicate better text-image alignment.

3. More Results

3.1. Text-Driven Appearance Generation

To demonstrate GAP's advantage over mesh-based methods with reconstructed geometries, we provide extensive visual comparisons in Fig. 2. We compare our results with both traditional geometry-based reconstruction using Ball-Pivoting Algorithm (BPA) [1] and learning-based reconstruction using CAPUDF [8]. For each reconstruction method, we generate UV maps using xatlas and apply the same texture generation methods (Texture, Text2Tex, Paint3D, SyncMVD) as baselines.

The visual comparison reveals two major challenges when using reconstructed meshes. First, both BPA [1] and CAPUDF [8] reconstructed meshes suffer from geometric ambiguities and information loss during surface reconstruction. Second, the excessive number of faces in reconstructed meshes leads to highly fragmented UV layouts with severe stretching and overlapping issues, as shown in Fig. 1. These fragmented UV charts not only limit the effective texture resolution but also cause color bleeding artifacts across chart boundaries, resulting in discontinuities and inconsistencies in the final appearance.

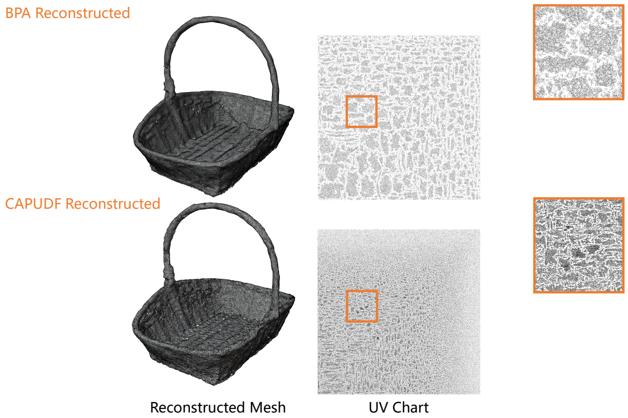

Figure 1. Comparison of UV parameterization results for reconstructed meshes using BPA and CAPUDF. The UV layouts exhibit severe fragmentation, stretching, and overlapping issues.

In contrast, GAP directly optimizes Gaussian primitives in 3D space without requiring intermediate mesh reconstruction or UV mapping. This direct optimization approach preserves geometric details from the input point cloud while enabling high-quality appearance generation across different object categories.

3.2. More Application Results

We further show more visualizations and comparisons of the application shown in the main Paper. We show more comparisons on the task of Point-to-Gaussian generation in Fig. 3. More visualizations on the Gaussian generations under real-world scanned DeepFashion3D dataset is shown in Fig. 4. Finally, we show more comparisons on learning to generate appearances for real-world 3D scenes in Fig. 5.

References

[1] Fausto Bernardini, Joshua Mittleman, Holly Rushmeier, Cláudio Silva, and Gabriel Taubin. The ball-pivoting algorithm for surface reconstruction. *IEEE Transactions on Visualization and Computer Graphics*, 5(4):349–359, 1999. 2

[2] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. In *International Conference on Learning Representations*, 2018. 1

[3] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning Transferable Visual Models From Natural Language Supervision . In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1, 2

[4] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1

[5] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016. 1

[6] Yu Yu, Weibin Zhang, and Yun Deng. Frechet inception distance (fid) for evaluating gans. *China University of Mining Technology Beijing Graduate School*, 3, 2021. 1

[7] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *IEEE International Conference on Computer Vision (ICCV)*, 2023. 1

[8] Junsheng Zhou, Baorui Ma, Yu-Shen Liu, Yi Fang, and Zhizhong Han. Learning Consistency-Aware Unsigned Distance Functions Progressively from Raw Point Clouds. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2

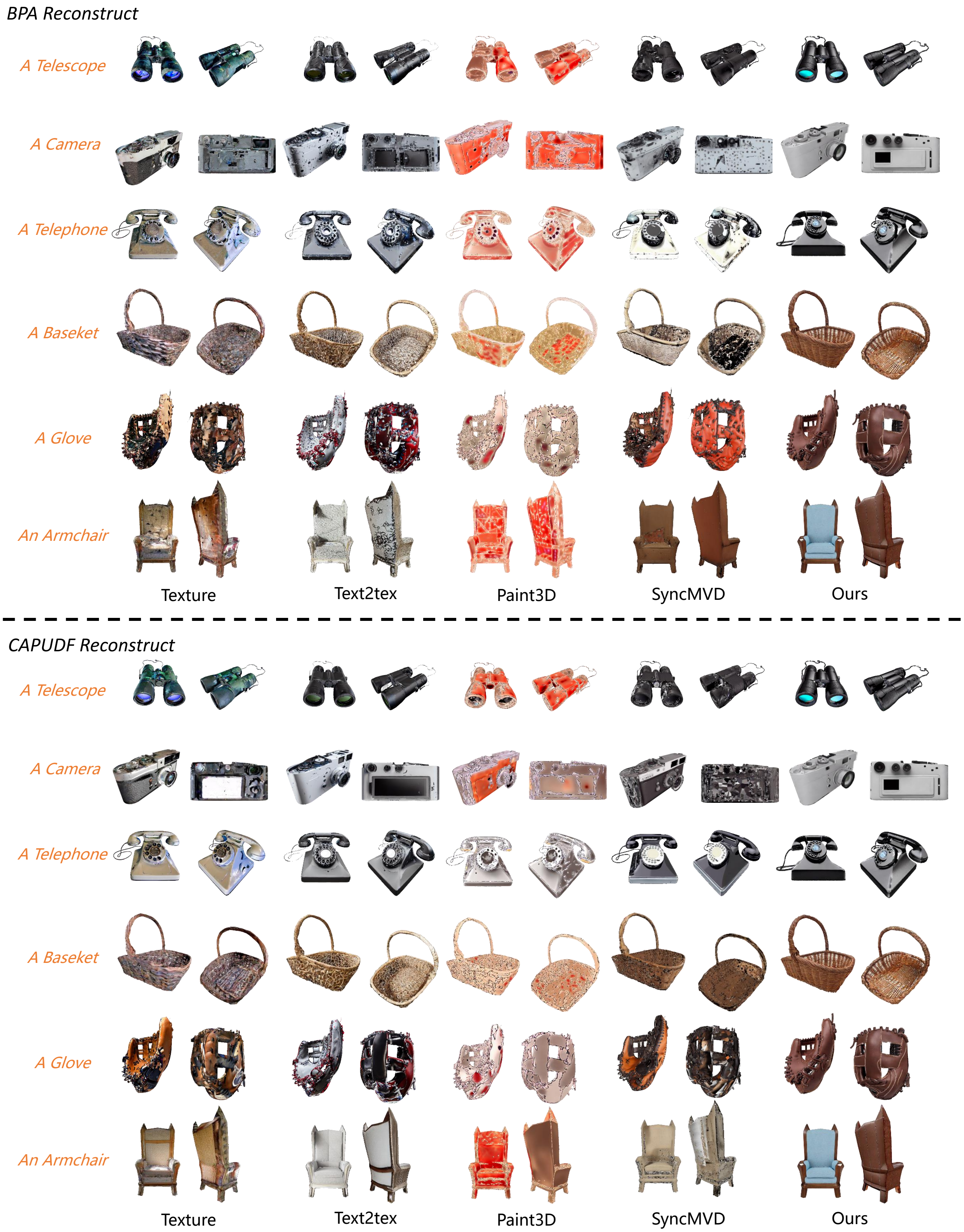

Figure 2. Visual comparison of text-guided appearance generation results with the reconstructed meshes.

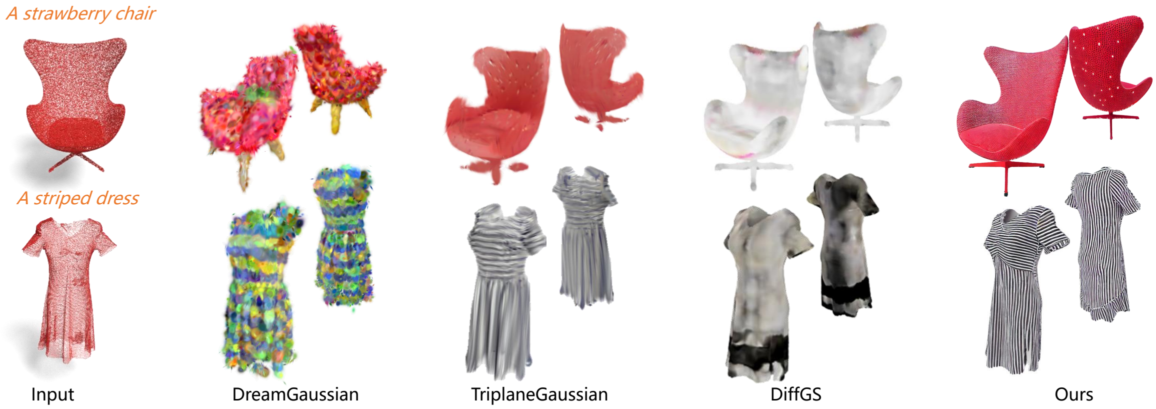

Figure 3. Visual comparison of point-to-Gaussian generation results on ShapeNet Chair and DeepFashion3D. Our GAP method demonstrates superior visual quality and geometric accuracy with flexible text-guided appearance control.

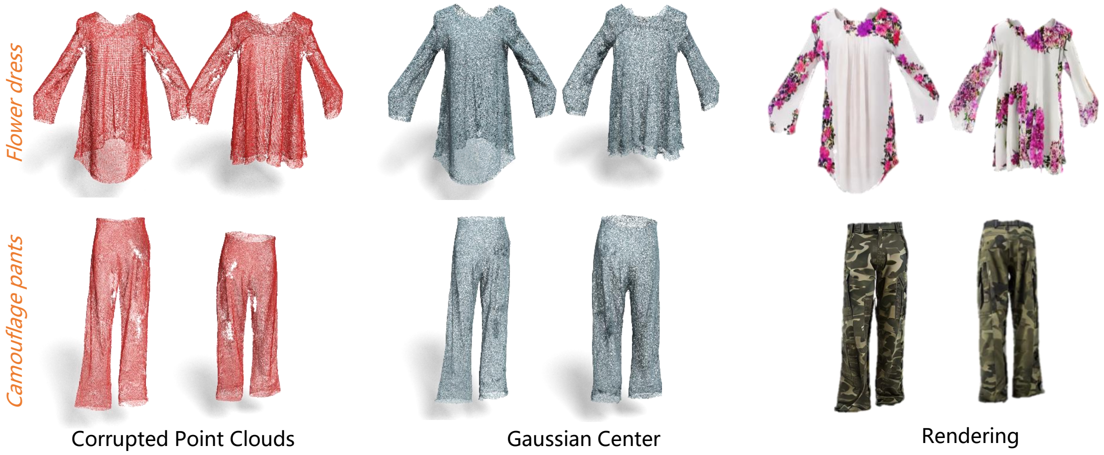

Figure 4. Gaussianization results on real-world partial scans from SRB and DeepFashion3D datasets. Our surface-anchoring mechanism and diffusion-guided rendering supervision enable GAP to generate complete, high-quality 3D Gaussian representations while maintaining geometric consistency.

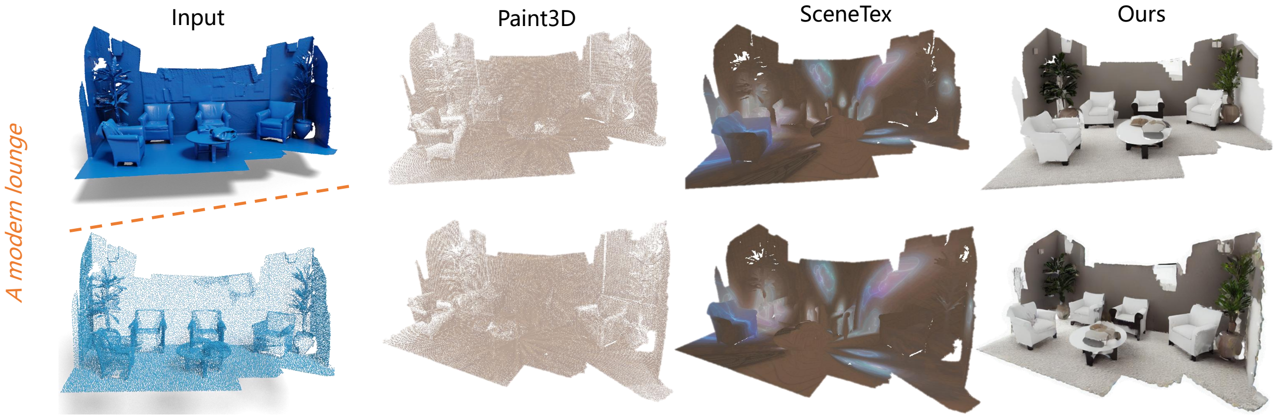

Figure 5. Scene-level Gaussianization comparison on real-world scanned 3DScene datasets. GAP generates high-quality results for complex scenes through a single optimization process.