CONFIDENTIAL EXTENDED ABSTRACT.
DO NOT DISTRIBUTE ANYWHERE.

カメラ位置未知少数視点からの3D Gaussian推定モデルの体系化と効果的なモデル構成による性能向上

藤村 友貴[1,a)]　櫛田 貴弘[2]　北野 和哉[1]　舩冨 卓哉[1]　向川 康博[1]

概要

本研究では、未知のカメラ位置で撮影された少数の画像から 3D Gaussian を推定するモデルについて、画像からの特徴抽出部、多視点間の情報集約部、出力部の 3 つのモジュールへの分割による体系化を提案する。この体系化に基づき、各モジュールにおける役割と設計、及び効果的なモデル構成についての議論を行う。本研究で提案するモデルは従来手法に対して、新規視点合成の精度向上と学習時に要する GPU メモリの約 25%の削減を達成した。本研究における計算効率の改善は、個人の研究者が最新の技術を実装する一助になることが期待される。

1. はじめに

近年コンピュータビジョン分野では、Neural radiance field (NeRF) [13] や 3D Gaussian splatting (3DGS) [11] といった、多視点画像を入力とした新規視点合成の研究が盛んに行われている。一方で、これらの手法の問題として、既知のカメラ位置から撮影された多数の画像を必要とすることが挙げられる。視点が少数の場合は新規視点合成のための観測が不足するだけでなく、カメラ位置の推定そのものも困難となる。

観測が不足する問題に対するアプローチの一つとして、データセットで feed-forward なモデルを学習することが挙げられる [2], [4], [5], [6], [24]。Large reconstruction model (LRM) [10] は、大規模なデータセットで学習することで 1 枚の画像からの NeRF の推定を可能としたモデルである。LRM を複数視点へ応用した研究 [25] や、カメラ位置が未知の問題設定へ拡張した研究 [17], [22] も存在する。また近年は、カメラ位置未知の 2 視点から 3 次元点群を推定する大規模モデル（DUSt3R [18]、MASt3R [16]）が提案されており、推定された 3 次元点群は 3DGS の初期値として用いることができる [9]。さらに、3 次元点群に加えて 3DGS のパラメータを直接出力するように拡張することで、カメラ位置未知の 2 視点から新規視点合成を可能とする技術も提案されている [15], [23]。

本研究は、カメラ位置が未知の少数視点から 3DGS のパラメータを推定することを目的とする。しかしながら、前述したようにわずか数ヶ月のうちに多くの研究が提案されており、また近年提案されるモデルは大規模化が進んでおり、手法の再現には強力な計算資源を要する。そこで本研究では従来手法に対して、**画像からの特徴抽出部、多視点間の情報集約部、出力部**の 3 つのモジュールへのモデル構成の分割による体系化を提案する。この体系化に基づき、各モジュールにおける役割及び設計について議論を行い、精度と計算効率の両面において効果的なモデルの提案を行う。実験では、提案したモデルが新規視点合成の精度を向上しつつ、学習時に要する GPU メモリの削減が可能であることを示す。本研究における計算効率の改善は、個人の研究者が最新の技術を実装する一助になることが期待される。

2. 3D Gaussian 推定モデル

2.1 モデル構成の体系化

本研究では未知のカメラ位置で撮影された少数の画像を入力とする。ここで、$\{I_i\}_{i=1,\cdots,N}$ を入力画像とし、N を入力枚数とする。各画像は $I_i \in \mathbb{R}^{H\times W\times 3}$ である。3D Gaussian を推定するモデルは以下のように記述できる。

$$\{T'_1,\cdots,T'_N\} = \mathcal{A}(\{\mathcal{E}(I_1),\cdots,\mathcal{E}(I_N)\}) \tag{1}$$

$$G_i = \mathcal{D}(T'_i) \quad \forall i \in \{1,\cdots,N\} \tag{2}$$

ここで、T'_i は 2.3 で説明する各画像の特徴トークンである。$G_i = \{\mu_i, \alpha_i, r_i, s_i, c_i\}$ は各画像のピクセルに対応した 3D Gaussian のパラメータであり、平均 $\mu_i \in \mathbb{R}^{H\times W\times 3}$、透過率 $\alpha_i \in \mathbb{R}^{H\times W\times 1}$、回転 $r_i \in \mathbb{R}^{H\times W\times 4}$、スケール $s_i \in \mathbb{R}^{H\times W\times 3}$、球面調和関数の係数 $c_i \in \mathbb{R}^{H\times W\times k}$ からなる。また、$\mathcal{E}$ は**画像からの特徴抽出部**、$\mathcal{A}$ は**多視点間の情報集約部**、$\mathcal{D}$ は**出力部**である。本研究では従来手法の各モジュールをこの分割へ割当て、表 1 に示すような体系化を提案する。モデル構成の他に、各手法の入力枚数、カメラ位置の有無、そして出力形式（点群 $P \in \mathbb{R}^{H\times W\times 3}$、NeRF

[1] 奈良先端科学技術大学院大学
[2] 立命館大学
[a)] fujimura.yuki@is.naist.jp

表 1 モデルの体系化

モデル	入力	カメラ位置	$\mathcal{E}$	$\mathcal{A}$	$\mathcal{D}$	出力
DUSt3R [18], MASt3R [16]	2	未知	Pretrained CroCo	2 視点間の cross attention	DPT head	P
Splatt3R [15], NoPoSplat [23]	2	未知	Pretrained MASt3R $\mathcal{E}$	2 視点間の cross attention	DPT head	G
LRM [10]	1	—	Convolution	（1 視点内での self-attention）	Deconvolution	θ
PF-LRM [17]	N	未知	Pretrained DINO	全視点間の self-attention	Deconvolution	θ
GS-LRM [25]	N	既知	Convolution	全視点間の self-attention	Linear+reshape	G
FreeSplatter [22]	N	未知	Convolution	全視点間の self-attention	Linear+reshape	G
提案モデル	N	未知	Pretrained MASt3R $\mathcal{E}$	全視点間の self-attention	DPT head	G

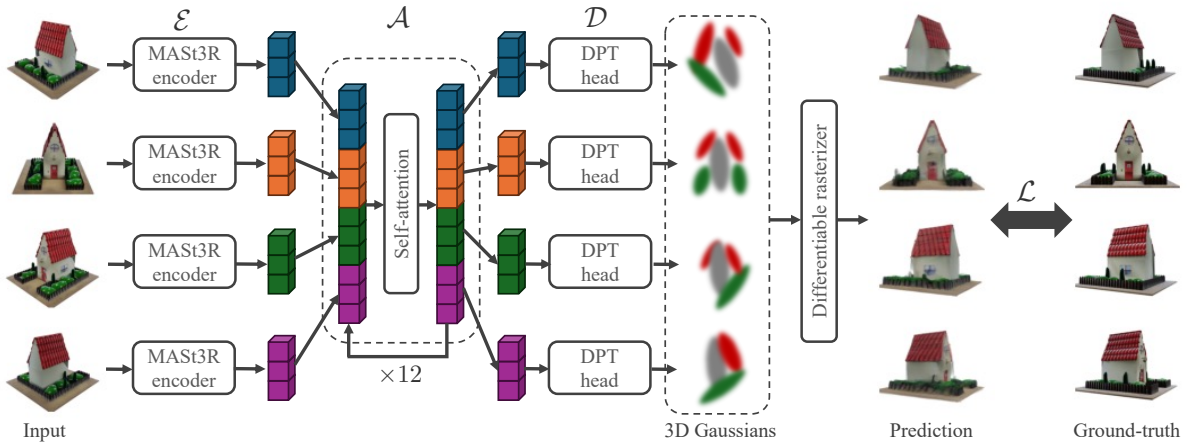


図 1 3D Gaussian 推定モデル

のパラメータ θ、Gaussian G）も記載している。

2.2 画像からの特徴抽出（$\mathcal{E}$）

まず最初に各入力画像に対して独立に特徴抽出を行う。

$$T_i = \mathcal{E}(I_i) \tag{3}$$

ここで、T_i は画像内のパッチごとの特徴トークンを表す。パッチサイズを $p \times p$、各パッチにおける特徴量の次元を d としたとき、T_i は $T_i \in \mathbb{R}^{\frac{H}{p} \times \frac{W}{p} \times d}$ となり、トークン数は HW/p^2 個である。

DUSt3R [18] と MASt3R [16] では CroCo [19], [20] を $\mathcal{E}$ として用いている。CroCo は Vision Transformer (ViT) [8] をベースとしたモデルであり、幾何情報の事前知識を 2 視点間の補完問題として自己教師あり学習したものである。Splatt3R [15] と NoPoSplat [23] は、MASt3R を 3D Gaussian のパラメータを出力するように拡張しており、したがって学習済み MASt3R の $\mathcal{E}$ を初期値として利用する。これらに対し PF-LRM [17] では、事前学習された大規模モデルである DINO [1] を使用している。一方で、GS-LRM [25] や FreeSplatter [22] では、ストライド p の畳み込み層 1 層のみでトークンを生成している。そのため、2.3 で説明する多視点間の情報集約部においても、画像特徴を学習するために多数の層を必要とする。

2.3 多視点間の情報集約（$\mathcal{A}$）

このモジュールでは、$\mathcal{E}$ で得られた各視点のトークン間で情報を集約する。

$$\{T'_1, \cdots, T'_N\} = \mathcal{A}(\{T_1, \cdots, T_N\}) \tag{4}$$

ここで、$T'_i \in \mathbb{R}^{\frac{H}{p} \times \frac{W}{p} \times d}$ は集約後の各トークンである。

DUSt3R と MASt3R は入力が 2 視点であるため、このモジュールでは各視点で別々の ViT を用意し、各視点間で cross attention を行うことで情報集約を実現している。一方で、PF-LRM、GS-LRM, FreeSplatter では $N > 2$ の視点を入力とするため、各視点のトークンを全て結合した NHW/p^2 個のトークン間で self-attention を行うことで情報集約を実現している。しかしながら、この方法は NHW/p^2 の二乗の attention を計算する必要があるため、計算コストが非常に大きい。

2.4 出力（$\mathcal{D}$）

最後に多視点の情報を集約した特徴トークンを用いて、各視点それぞれで出力を行う。3D Gaussian を出力する場合は、

$$G_i = \mathcal{D}(T'_i) \tag{5}$$

として、各画像の各ピクセルに対応する 3D Gaussian のパラメータを推定する。

トークンからピクセルごとの密な推定をする必要があ

るため、DUSt3R 等では DPT head [14] を用いる。DPT head は複数の畳み込み、逆畳み込み層で構成されており、トークンを元の解像度に復元することで密な推定を可能としている。一方で、FreeSplatter 等では各 d 次元トークンを線形層 1 層で p^2K 次元にしたあと reshape することで各パッチごとに $p \times p \times K$ の密な推定を行う。ここで、K は出力対象の次元であり、3D Gaussian を出力する場合は 2.1 で述べたパラメータの次元を全て足し合わせた $3+1+4+3+k = 11+k$ となる。この方法はトークンから線形層 1 層で元の解像度の復元を行うため、実験で示すように出力結果に格子状のアーティファクトが生じる。

2.5 効果的なモデル構成

前節で全体のモデルの 3 つのモジュールへの分割と、各モジュールの役割について議論を行なった。本研究は未知のカメラ位置で撮影された多視点画像を入力とした 3D Gaussian の推定を目的とし、前節の議論に基づいて以下の指針に沿って各モジュールの設計を行う。

- 特徴抽出部では幾何情報の強力な事前知識を有したモジュールを用い、精度の向上を図る。
- 情報集約部では全視点トークン間で self-attention を行うため、モジュールを軽量化する。
- 出力部では密な推定を可能とするモジュールを用いる。

表 1 下部と図 1 に提案する具体的なモデルを示す。特徴抽出部では、Splatt3R と NoPoSplat 同様事前学習済みの MASt3R の $\mathcal{E}$ を初期値として用いる。出力部では DPT head を用いることで密な推定を可能とする。

情報集約部では、多視点間の情報集約を行うため全視点間の self-attention を行う。GS-LRM と FreeSplatter では特徴抽出部が畳み込み層 1 層のため、このモジュールの self-attention ブロックの数は 24 である。また、出力部についても線形層 1 層であり、密な出力のためにトークンのパッチサイズが 8 と小さく、計算コストが非常に高い。一方で、本研究では特徴抽出部に事前学習された MASt3R の $\mathcal{E}$ を、加えて出力部に DPT head を用いるため、self-attention のブロック数とトークンのパッチサイズをそれぞれ 12 と 16 に設定した。これにより、精度を低下させることなく大幅な計算コストの削減が可能である。

2.6 学習

図 1 に示すように、学習時は多視点画像 $\{I_i\}_{i=1,\cdots,N}$ を入力し、3D Gaussian $\{G_i\}_{i=1,\cdots,N}$ を推定する。その後、推定した 3D Gaussian を用いて入力した画像とは別視点の画像をレンダリングし、正解の画像と比較することで学習を行う。なお、学習時は別視点のレンダリングのため正解のカメラ位置を用いる。

具体的な誤差関数は以下である。

表 2 実験での比較モデル（$\mathcal{A}$ は (ブロック数, パッチサイズ)）

モデル	$\mathcal{E}$	$\mathcal{A}$	$\mathcal{D}$
FreeSplatter	Convolution	(24,8)	Linear+reshape
Ours1	MASt3R $\mathcal{E}$	(12,16)	Linear+reshape
Ours2	MASt3R $\mathcal{E}$	(12,16)	DPT head

表 3 新規視点合成の定量評価とパラメータ数・必要 GPU メモリ数

モデル	PSNR	SSIM	LPIPS	Params.	GPU [GB]
FreeSplatter	21.1	0.846	0.169	307M	225
Ours1	**23.4**	0.877	0.114	461M	172
Ours2	**23.4**	**0.879**	**0.111**	475M	174

$$\mathcal{L} = \mathcal{L}_m + \lambda_{\mathrm{l}}\mathcal{L}_l + \lambda_{\mathrm{3D}}\mathcal{L}_{\mathrm{3D}} \tag{6}$$

ここで、$\mathcal{L}_m$ はレンダリングした画像と正解画像の平均二乗誤差、$\mathcal{L}_l$ は画像間の知覚的損失 [26] である。後述するように、本研究では学習に合成画像のみを用いるため、正解の深度画像及び 3 次元点群が利用可能である。したがって FreeSplatter 同様、推定した 3D Gaussian の平均と正解の 3 次元点群との二乗誤差 $\mathcal{L}_{\mathrm{3D}}$ も学習に用いた。なお、本研究では $\lambda_l = 0.05$、$\lambda_{\mathrm{3D}} = 10$ とした。

3. 実験

提案したモデルの有効性を検証するため、表 2 に示すモデル間で比較を行なった。ベースラインを FreeSplatter [22] とし、Ours1 では特徴抽出部を学習済み MASt3R [16] の $\mathcal{E}$ に置き換え、情報集約部の self-attention のブロック数とパッチサイズをそれぞれ 12、16 に設定した。Ours2 ではさらに出力部を DPT head [14] に置き換えた。

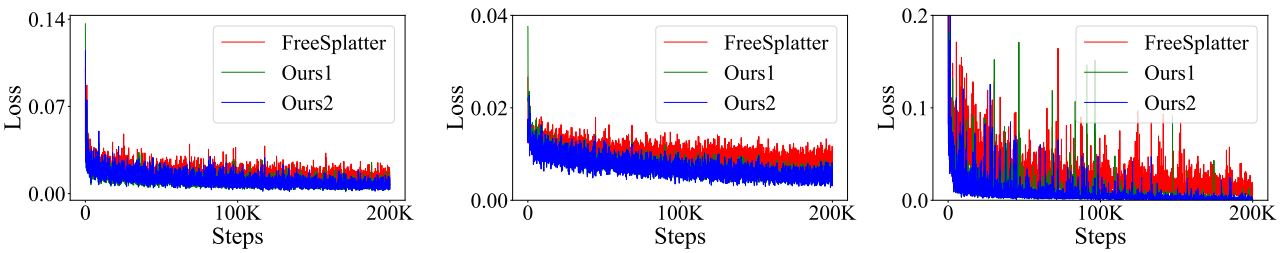
3.1 データセット

学習・評価用データセットとして G-Objaverse [21] データセットを使用した。このデータセットは 3D オブジェクトのデータセット Objaverse [7] をレンダリングしたものである。264,775 シーンからなり、学習と評価のため 10 : 1 に分割を行なった。各シーンは物体を囲む 38 視点からの画像が含まれている。LaRa [3] と同様、38 視点のカメラ位置を K-means を用いて 4 クラスタにクラスタリングし、学習時は各クラスタから 1 枚ずつ、計 4 枚の画像を選択しモデルに入力する。誤差関数の計算のための新規視点についても、同様の方法で 4 枚選択する。なお、入力画像の解像度は 256×256 とした。

3.2 実装

最適化手法は AdamW [12] を使用した。学習率は 5.0×10^{-5} で、最初の 2000 ステップの Linear warmup ののち、 Cosine annealing により学習終了までに 5.0×10^{-6} まで減衰させた。バッチサイズは 32、学習ステップ数は 200K である。$\mathcal{A}$ と $\mathcal{D}$ については bf16 の混合精度で学習し

図 2　新規視点合成と 3D Gaussian の depth の比較



(a) $\mathcal{L}_m$　(b) $\mathcal{L}_l$　(c) $\mathcal{L}_{3D}$

図 3　学習中の各誤差関数の値のプロット

た。計算環境は NVIDIA A100 GPU (80 GB) ×4 である。

3.3　実験結果

新規視点合成

表 3 に新規視点合成の定量評価を示す。FreeSplatter と比較して Ours1 と Ours2 のどちらも精度が大幅に向上しており、MASt3R の特徴抽出部の利用が重要であることがわかる。図 2 は定性評価であり、(a) は入力に用いた画像、(b–e) はそれぞれ新規視点の正解画像と推定画像である。(f,g) は深度画像で、1 枚目の入力画像（(a) 左上）から出力された 3D Gaussian の z 座標を可視化したものである。なお、深度画像は正解画像でマスクしてある。Ours2 は DPT head により密な推定が可能であるため、Ours1 の深度画像における格子状のアーティファクトが軽減され、新規視点合成においても画像の細部の再現が可能である。

学習の効率化

表 3 に各モデルの学習パラメータ数と学習に要する GPU メモリを示す。Ours1、Ours2 は FreeSplatter より学習パラメータ数が多いものの、少ないブロック数と大きいパッチサイズの self-attention ブロックを用いることにより、約 25%の GPU メモリの削減が可能である。

図 3 に学習中の各誤差関数の値のプロットを示す。Ours1、Ours2 は FreeSplatter に比べて学習誤差が大幅に減少している。各ステップの計算量も減少しており、FreeSplatter は 200K ステップに約 4 日要するのに対し Ours2 は 3 日ほどであり、約 25%の学習速度の改善が見られた。

4.　まとめ

本研究では、未知のカメラ位置から撮影された少数の画像から 3D Gaussian を推定するモデルについて、従来手法の体系化を提案し、それに基づいて精度・計算効率の両面で効果的なモデル構成を提案した。近年提案されるモデルは大規模化が進んでおり、手法の再現には強力な計算資源を要する。本研究で提案したモデルの計算効率の改善は、個人の研究者が最新の技術を実装する一助になることが期待される。

謝辞

本研究は、JST ムーンショット型研究開発事業 JPMJMS2389 の支援を受けた。また、産総研及び AIST Solutions が提供する ABCI 3.0 を「ABCI 3.0 開発加速利用」を支援を受けて利用した。

参考文献

[1] Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P. and Joulin, A.: Emerging Properties in Self-Supervised Vision Transformers, *ICCV*, pp. 9650–9660 (2021).

[2] Charatan, D., Li, S. L., Tagliasacchi, A. and Sitzmann, V.: pixelSplat: 3D Gaussian Splats from Image Pairs for Scalable Generalizable 3D Reconstruction, *CVPR*, pp. 19457–19467 (2024).

[3] Chen, A., Xu, H., Esposito, S., Tang, S. and Geiger, A.: LaRa: Efficient Large-Baseline Radiance Fields, *ECCV*, pp. 338–355 (2024).

[4] Chen, A., Xu, Z., Zhao, F., Zhang, X., Xiang, F., Yu, J. and Su, H.: MVSNeRF: Fast Generalizable Radiance Field Reconstruction From Multi-View Stereo, *ICCV*, pp. 14124–14133 (2021).

[5] Chen, Y., Xu, H., Zheng, C., Zhuang, B., Pollefeys, M., Geiger, A., Cham, T.-J. and Cai, J.: MVSplat: Efficient 3D Gaussian Splatting from Sparse Multi-view Images, *ECCV*, pp. 370–386 (2024).

[6] Chen, Y., Zheng, C., Xu, H., Zhuang, B., Vedaldi, A., Cham, T.-J. and Cai, J.: MVSplat360: Feed-Forward 360 Scene Synthesis from Sparse Views, *NeurIPS* (2024).

[7] Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A. and Farhadi, A.: Objaverse: A Universe of Annotated 3D Objects, *CVPR*, pp. 13142–13153 (2023).

[8] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. and Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, *ICLR* (2021).

[9] Fan, Z., Wen, K., Cong, W., Wang, K., Zhang, J., Ding, X., Xu, D., Ivanovic, B., Pavone, M., Pavlakos, G., Wang, Z. and Wang, Y.: InstantSplat: Sparse-view SfM-free Gaussian Splatting in Seconds, *arXiv preprint arXiv:2403.20309* (2024).

[10] Hong, Y., Zhang, K., Gu, J., Bi, S., Zhou, Y., Liu, D., Liu, F., Sunkavalli, K., Bui, T. and Tan, H.: LRM: Large Reconstruction Model for Single Image to 3D, *ICLR* (2024).

[11] Kerbl, B., Kopanas, G., Leimkuehler, T. and Drettakis, G.: 3D Gaussian Splatting for Real-Time Radiance Field Rendering, *ACM TOG*, Vol. 42, No. 4 (2023).

[12] Loshchilov, I. and Hutter, F.: Decoupled Weight Decay Regularization, *ICLR* (2019).

[13] Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R. and Ng, R.: NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis, *ECCV* (2020).

[14] Ranftl, R., Bochkovskiy, A. and Koltun, V.: Vision Transformers for Dense Prediction, *ICCV* (2021).

[15] Smart, B., Zheng, C., Laina, I. and Prisacariu, V. A.: Splatt3R: Zero-shot Gaussian Splatting from Uncalibrated Image Pairs, *arXiv preprint arXiv:2408.13912* (2024).

[16] Vincent Leroy, Yohann Cabon, J. R.: Grounding Image Matching in 3D with MASt3R, *arXiv preprint arXiv:2406.09756* (2024).

[17] Wang, P., Tan, H., Bi, S., Xu, Y., Luan, F., Sunkavalli, K., Wang, W., Xu, Z. and Zhang, K.: PF-LRM: Pose-Free Large Reconstruction Model for Joint Pose and Shape Prediction, *ICLR* (2024).

[18] Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B. and Revaud, J.: DUSt3R: Geometric 3D Vision Made Easy, *CVPR*, pp. 20697–20709 (2024).

[19] Weinzaepfel, P., Lucas, T., Leroy, V., Cabon, Y., Arora, V., Brégier, R., Csurka, G., Antsfeld, L., Chidlovskii, B. and Revaud, J.: CroCo v2: Improved Cross-view Completion Pre-training for Stereo Matching and Optical Flow, *ICCV* (2023).

[20] Weinzaepfel, Philippe and Leroy, Vincent and Lucas, Thomas and Brégier, Romain and Cabon, Yohann and Arora, Vaibhav and Antsfeld, Leonid and Chidlovskii, Boris and Csurka, Gabriela and Revaud Jérôme: CroCo: Self-Supervised Pre-training for 3D Vision Tasks by Cross-View Completion, *NeurIPS* (2022).

[21] Xu, C., Dong, Y., Zuo, Q., Zhang, J., Ye, X., Geng, W., Zhang, Y., Gu, X., Qiu, L., Zhao, Z., Ran, Q., Jiang, J., Dong, Z. and Bo, L.: G-buffer Objaverse: High-Quality Rendering Dataset of Objaverse.

[22] Xu, J., Gao, S. and Shan, Y.: FreeSplatter: Pose-free Gaussian Splatting for Sparse-view 3D Reconstruction, *arXiv preprint arXiv:2412.09573* (2024).

[23] Ye, B., Liu, S., Xu, H., Li, X., Pollefeys, M., Yang, M.-H. and Peng, S.: No Pose, No Problem: Surprisingly Simple 3D Gaussian Splats from Sparse Unposed Images, *ICLR* (2025).

[24] Yu, A., Ye, V., Tancik, M. and Kanazawa, A.: pixelNeRF: Neural Radiance Fields From One or Few Images, *CVPR*, pp. 4578–4587 (2021).

[25] Zhang, K., Bi, S., Tan, H., Xiangli, Y., Zhao, N., Sunkavalli, K. and Xu, Z.: GS-LRM: Large Reconstruction Model for 3D Gaussian Splatting, *ECCV*, pp. 1–19 (2024).

[26] Zhang, R., Isola, P., Efros, A. A., Shechtman, E. and Wang, O.: The Unreasonable Effectiveness of Deep Features as a Perceptual Metric, *CVPR* (2018).