少数の in-the-wild 画像からの 3D ガウシアン推定モデル

藤村 友貴[1,a)]　櫛田 貴弘[2]　北野 和哉[1]　舩冨 卓哉[3]　向川 康博[1]

概要

本研究では、カメラパラメータが未知である少数の in-the-wild 画像から、3D ガウシアンを feed-forward で推定するモデルを提案する。In-the-wild 画像とは、例えばインターネット上にあるような、異なるカメラ、日時に撮影された画像である。本研究では既存の 3D ガウシアン推定モデルを拡張し、3D ガウシアンと同時に照明環境を表現する埋め込みを学習する。この埋め込みを用いて 3D ガウシアンの色を変化させることで、in-the-wild 画像における大きな照明変化を表現することが可能になる。実験では、たった 2 枚の in-the-wild 画像から、新規視点生成と照明環境のコントロールが可能であることを示す。

1. はじめに

多視点画像からの 3 次元再構成や新規視点生成はコンピュータビジョン分野における主要なテーマの一つである。3D Gaussian splatting (3DGS) [4] は対象を色付きの 3D ガウシアンの集合として表現し、入力した多視点画像を表現するように各ガウシアンのパラメータを学習することで、学習時に使用していない新規視点における画像の生成を可能とする技術である。3DGS も含めた一般の 3 次元再構成や新規視点生成は、カメラパラメータが既知、あるいは十分な精度で推定可能な数十から数百の視点数を必要とする。また、高精度なカメラパラメータ推定や 3 次元再構成には、同じ照明環境下で撮影された画像を用いることが望ましい。

これに対して本研究では、**「カメラパラメータが未知」**である**「少数」**の、そして**「照明などの撮影環境が異なる」**画像から 3D ガウシアンを推定する手法を提案する。具体的には in-the-wild 画像からの 3D ガウシアン推定モデルを提案する。In-the-wild 画像とは、例えばインターネット上にあるような、異なるカメラ、日時に撮影された画像である。提案手法ではこのような入力から feed-forward で 3D ガウシアンを推定する。また、3D ガウシアンと同時に照明環境を表現する埋め込みも学習を行う。図 1 に示すように、提案手法では照明環境の異なるたった 2 枚の画像から、新規視点生成と照明環境のコントロールが可能である。

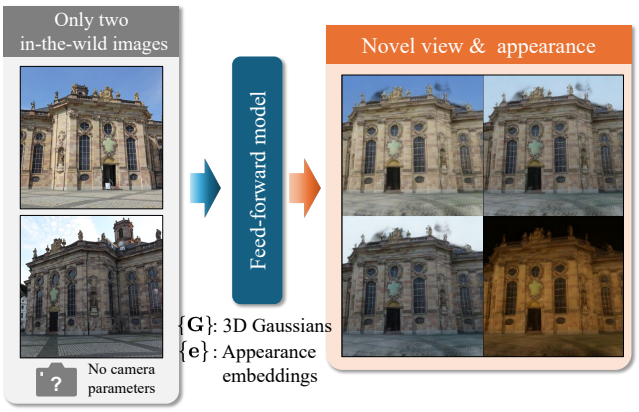

図 1　2 枚のカメラパラメータが未知である in-the-wild 画像を入力し、3D ガウシアンと照明環境を表現する埋め込みを出力し、新規視点生成と照明環境のコントロールを可能とする。

2. 関連研究

2.1 In-the-wild 画像からの新規視点画像生成

NeRF-W [8] は in-the-wild 画像を用いた新規視点画像生成手法である。Neural radiance field (NeRF) [9] によるシーン表現に加え、画像ごとに照明環境を表現する埋め込みを同時に学習し、in-the-wild 画像からの新規視点生成と照明環境のコントロールを可能とする。WildGaussians [5] は in-the-wild 画像を用いて 3DGS を学習する手法であり、画像と各ガウシアンごとに照明環境を表現する埋め込みを学習する。しかしながら、これらの手法は学習に数百枚以上の画像を必要とし、またそれらのカメラパラメータも既知である必要がある。近年は数枚から数十枚で動作する手法が提案されているものの [6], [16]、依然カメラパラメータを必要とする。さらに、これらの手法は NeRF や 3D ガウシアンをシーンごとに学習する必要もあり、手法によっては学習に数時間以上を要する。

2.2 Feed-forward な 3D ガウシアン推定モデル

近年、入力画像から feed-forward で 3D ガウシアンを推定するニューラルネットワークモデルが多く提案されてい

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

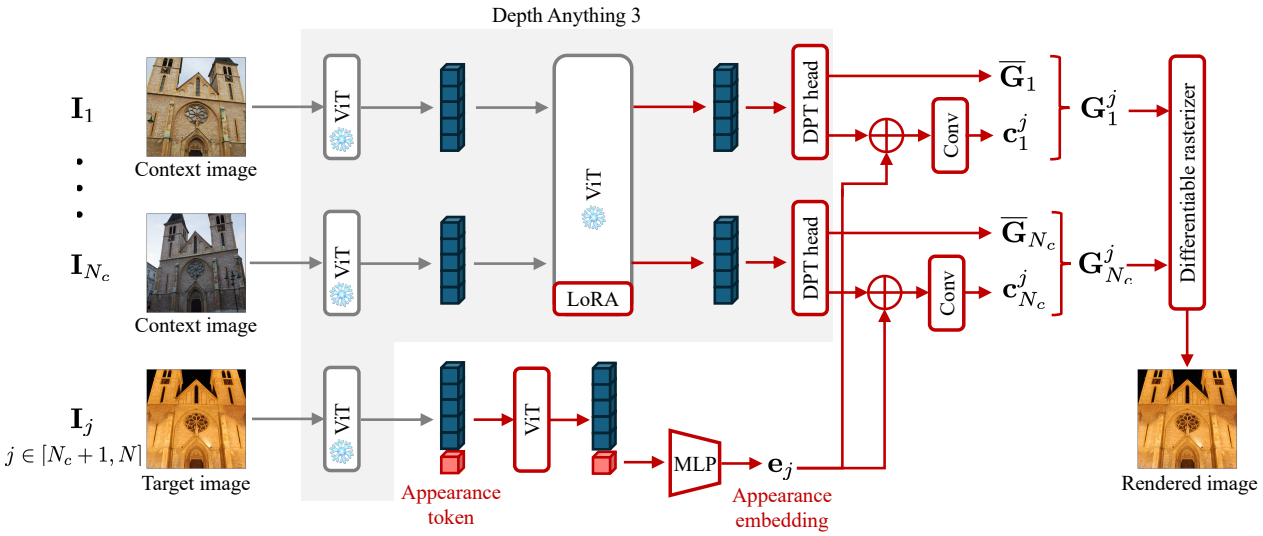

図 2　提案手法の全体像

る [1], [2], [15]。Feed-forward なモデルはシーンごとの最適化を必要とせず、数秒ほどで 3D ガウシアンの推定が可能である。また、シーンに対する事前知識を学習することで、数枚の画像から 3D ガウシアンを推定することができる。さらに近年は、カメラパラメータが不要である手法も多く提案されている [7], [13], [14], [18]。しかしながら、これらの手法は in-the-wild 画像の入力を想定しておらず、出力された 3D ガウシアンは画像間で大きく異なる照明環境を十分に表現することができない。

3. 提案手法

3.1 概要

提案手法の全体像を図 2 に示す。入力は $N = N_c + N_t$ 枚の in-the-wild 画像 $\{\mathbf{I}_i\}_{i=1}^{N}$ である。各画像は $\mathbf{I}_i \in \mathbb{R}^{H\times W\times 3}$ で画像サイズは $H \times W$ である。N_c 枚の画像については、各画像に対応する 3D ガウシアン $\{\mathbf{G}\}_{i=1}^{N_c}$ を出力する。学習時はこれらの 3D ガウシアンから、ターゲットとなる N_t 枚の画像をレンダリングし、入力として与えた画像との誤差を計算しモデルの学習を行う。

3.2 モデルアーキテクチャ

提案手法は Depth Anything 3 (DA3) [7] をベースとする。まず初めに、全ての画像を独立に Vision Transformer (ViT) に入力し特徴トークンを取得する。その後、N_c 枚の画像については、全画像間での self-attention を含む ViT に入力し、画像間の情報を集約する。最後に各画像ごとに DPT head [10] に入力し、3D ガウシアンを出力する。

各画像の 3D ガウシアンの位置 $\boldsymbol{\mu}_i \in \mathbb{R}^{H\times W\times 3}$ は、深度画像 $\mathbf{D}_i \in \mathbb{R}^{H\times W\times 1}$ と、カメラ中心と光線の方向からなる光線マップ $[\mathbf{o}_i, \mathbf{d}_i] \in \mathbb{R}^{H\times W\times 6}$ を用いて、$\boldsymbol{\mu}_i = \mathbf{o}_i + \mathbf{D}_i\mathbf{d}_i$ として計算される。後述するように本研究では、照明環境の違いによって 3D ガウシアンの色を表す球面調和関数の係数の値を変化させるため、不透明度 $\boldsymbol{\alpha}_i \in \mathbb{R}^{H\times W\times 1}$、回転 $\mathbf{r}_i \in \mathbb{R}^{H\times W\times 4}$、スケール $\mathbf{s}_i \in \mathbb{R}^{H\times W\times 3}$ のみをもつ $\{\overline{\mathbf{G}}_i\}_{i=1}^{N_c}$ を出力するようにする。

3.3 照明環境を表現する埋め込みの学習

3DGS では各 3D ガウシアンの色は視線方向に依存する球面調和関数の係数として表現される。しかしながら、in-the-wild 画像のような視点によって大きく照明環境が異なる画像では、球面調和関数のみでは視点依存の色を表現できない。そこで本研究では、照明環境を表す埋め込みを 3D ガウシアンと同時に学習し、球面調和関数の係数を埋め込みによって変化させる手法を提案する。

まず最初に、3D ガウシアンを推定する画像と同様に、学習時のターゲットとなる N_t 枚の画像をそれぞれ独立に ViT に入力する。その後、得られた特徴トークンに対して、学習可能なトークン（Appearance token（図 2））を追加し、これらを ViT に入力する。最後に出力されたトークンを multi-layer perceptron (MLP) に入力し、照明環境を表す 1 次元の埋め込みベクトル $\{\mathbf{e}_j\}_{j=N_c+1}^{N}$ を推定する。

学習時、j 番目（$j \in [N_c+1, N]$）の画像のレンダリングを行うとする。このとき、各画像の 3D ガウシアンを出力する DPT head の最終層から特徴マップを取り出し、埋め込みベクトル $\mathbf{e}_j$ を空間方向にコピーしてチャンネル方向に結合する。その後、一層の畳み込み層に入力し、球面調和関数の係数 $\mathbf{c}_i^j \in \mathbb{R}^{H\times W\times k}$ を出力する。最終的に $\{\overline{\mathbf{G}}_i\}_{i=1}^{N_c}$ と合わせることで、j 番目の画像の照明環境下である 3D ガウシアン $\{\mathbf{G}_i^j\}_{i=1}^{N_c}$ を得る。

3.4 損失関数

学習時は各ターゲット画像に対してそれぞれの照明環境

図 3 NeRF-OSR データセットでの定性評価

下での 3D ガウシアン $\{\{\mathbf{G}_i^j\}_{i=1}^{N_c}\}_{j=N_c+1}^{N}$ を生成し、ターゲット画像のレンダリングを行う。そして以下の損失関数を用いて学習を行う。

$$\mathcal{L} = \mathcal{L}_{\text{MSE}} + \lambda_{\text{lpips}}\mathcal{L}_{\text{lpips}} + \lambda_{\text{ray}}\mathcal{L}_{\text{ray}} \tag{1}$$

ここで、$\mathcal{L}_{\text{MSE}}$ と $\mathcal{L}_{\text{lpips}}$ はターゲット画像とレンダリングした画像間の平均二乗誤差と知覚的損失 [17] である。$\mathcal{L}_{\text{ray}}$ は光線マップに関する損失であり、カメラ位置の平均二乗誤差と光線方向のコサイン類似度からなる。

$$\mathcal{L}_{\text{ray}} = \sum_{i=1}^{N_c}\sum_{p\in\Omega} \|\mathbf{o}_i(p) - \hat{\mathbf{o}}_i(p)\|^2 + \frac{\mathbf{d}_i(p)^\top \hat{\mathbf{d}}_i(p)}{\|\mathbf{d}_i(p)\|\|\hat{\mathbf{d}}_i(p)\|} \tag{2}$$

ここで、$\Omega = [1, H] \times [1, W]$、$\hat{\mathbf{o}}_i$ と $\hat{\mathbf{d}}_i$ は入力した画像の正解のカメラ位置である。$\hat{\mathbf{o}}_i$ のスケールは学習時に $\mathbf{o}_i$ に合わせて計算を行う。

4. 実験

4.1 データセット

学習には MegaScenes データセット [12] を用いる。このデータセットには 458K シーンが含まれており、各シーンはインターネット上で収集された、異なる日時、カメラで撮影された画像が含まれる。また、各シーンには COLMAP で推定したカメラパラメータが含まれる。3D ガウシアンの学習を安定させるため、3749 シーンから画像間で遮蔽が少ない 10456 の多視点画像集合を抽出した。各画像集合は 3 視点からなり、したがって $[N_c, N_t] = [2, 1]$ とした。

評価には NeRF-OSR データセット [11] の 3 つのシーンを用いた。このデータセットには、異なる時間帯のそれぞれで画像列を撮影したものが含まれている。したがって、異なる時間と視点で撮影された 2 枚の画像を入力とし、それぞれと同じ時間帯で撮影された別視点の画像をレンダリングして評価を行った。評価画像は、それぞれの時間帯で 2 視点ずつ選択した。

表 1 NeRF-OSR データセットでの定量評価

Method	PSNR↑	LPIPS↓
DA3	14.445	0.470
Ours	**16.439**	**0.379**

4.2 実装

提案手法は学習済み DA3 をベースに実装した。前半の各画像を独立に処理する ViT のパラメータは固定し、後半の全画像間の情報を集約する ViT には LoRA [3] を用いた。その他の ViT と MLP、各ヘッドのパラメータはすべて学習を行った。損失関数の重みは $[\lambda_{\text{lpips}}, \lambda_{\text{ray}}] = [1.0, 0.1]$ とした。入力画像の解像度は 504×504 である。学習には NVIDIA A100 GPU×4 を使用した。

4.3 実験結果

図 3 に実験結果を示す。DA3 をベースラインとして比較を行った。

3 つのシーンのそれぞれについて、左から入力画像、DA3 と提案手法の新規視点生成、正解の画像を表示してある。各行は同じ時間帯に撮影された画像であり、提案手法では入力画像から得られた埋め込みを用いて、各環境下で 3D ガウシアンの色を変化させている。DA3 は照明環境が大きく異なる in-the-wild 画像の入力を想定しておらず、照明環境の違いを表現することが不可能である. 加えて、生成された画像に大きなノイズが含まれている。これに対して提案手法では、それぞれの環境下における色を再現することが可能である。

表 1 には PSNR と LPIPS を用いた定量評価を示しており、いずれも提案手法が DA3 を上回ることが確認された。

4.4 視点と照明環境の補間

提案手法は照明環境を 1 次元の埋め込みベクトルとして表現しているため、入力画像から得られた埋め込みベクトルの線形内挿を行うことで、視点の補間だけでなく照明環

図 4　1 行目：$\mathbf{I}_0$ から得られた照明環境の埋め込みで視点のみを補間。2 行目：$\mathbf{I}_0$ と $\mathbf{I}_1$ から得られた照明環境の埋め込みの線形内挿入により、視点と照明環境を補間。3 行目：$\mathbf{I}_1$ から得られた照明環境の埋め込みで視点のみを補間。

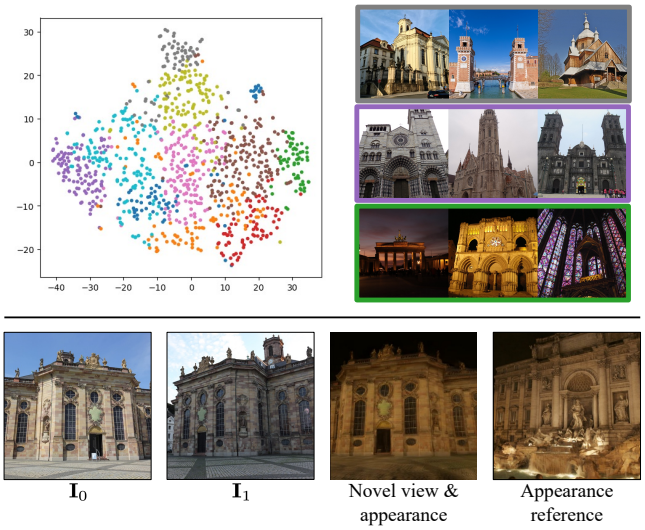

図 5　上段：照明環境の埋め込みの可視化と 3 つのクラスタ内の画像の例。下段：別のシーンから得られた埋め込みを用いて照明環境を変化させた例。

境の補間を行うことが可能である。図 4 は入力画像 $\mathbf{I}_0$ と $\mathbf{I}_1$ について、それぞれから推定した埋め込みで視点のみを補間したものと、埋め込みの線形内挿入により、視点と照明環境を補間した例である。

4.5　異なるシーンの画像を用いた照明環境の編集

図 5 上段は、学習画像から得られた照明環境の埋め込みを 2 次元に可視化したものである。また、3 つのクラスタに含まれる画像の例も表示している。異なるシーンにおいても照明環境が類似している場合は、得られた埋め込みが類似することがわかる。下段はこの性質を用いて、3D ガウシアンを推定するための入力画像 $\mathbf{I}_0$、$\mathbf{I}_1$ に対して、別のシーンの画像を用いて照明環境を変化させた例である。

4.6　一時的に存在する物体の学習

提案手法は in-the-wild 画像を用いて学習を行っているが、in-the-wild 画像では照明変化だけでなく、特定の視点のみに人や車といった一時的な物体が存在する可能性がある。本研究では学習時にこのような物体についてモデル化は行なっていないが、それらの影響を除去するようモデル自体が自動的に学習している可能性が示唆された。図 6 では、入力画像 $\mathbf{I}_0$、$\mathbf{I}_1$ について、ターゲットとなる画像とレンダリングした画像に加え、推定した 3D ガウシアンの不透明度を示している。$\mathbf{I}_1$ の赤枠内の物体は $\mathbf{I}_0$ とターゲット画像には存在しないが、推定した不透明度 $\boldsymbol{\alpha}_1$ の対応する箇所の値が 0 となっており、レンダリング時に無視されていることがわかる。

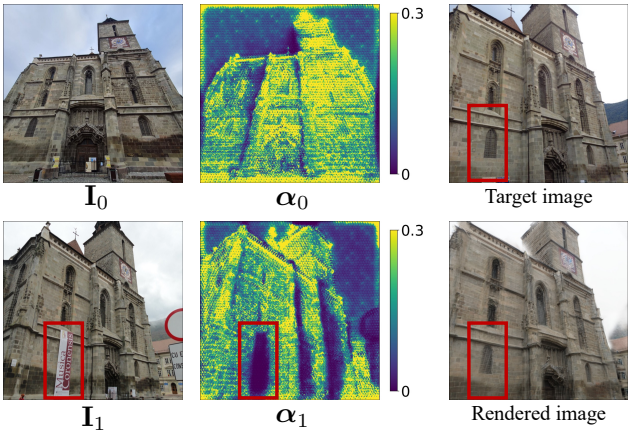

図 6　入力画像 $\mathbf{I}_0$、$\mathbf{I}_1$ について、$\mathbf{I}_1$ 中の赤枠内の物体は $\mathbf{I}_0$ に存在しないが、3D ガウシアンの不透明度 $\boldsymbol{\alpha}_1$ の対応箇所を 0 とすることで、レンダリング時に無視されるよう自動的に学習されている。

5.　まとめ

本研究では、少数のカメラパラメータが未知である in-the-wild 画像から、3D ガウシアンを feed-forward で推定するモデルを提案した。視点間の照明環境の差異に対して、照明環境を表現する埋め込みを同時に学習し、この埋め込みによって 3D ガウシアンの色を変化させる手法を提案した。実験では、学習した埋め込みが照明変化を再現し、視点と照明環境の補間が可能であることを示した。

謝辞　本研究は JSPS 科研費 26K02931 の助成を受けたものです。

参考文献

[1] 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).

[2] 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).

[3] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L. and Chen, W.: LoRA: Low-Rank Adaptation of Large Language Models, *ICLR* (2022).

[4] 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).

[5] Kulhanek, J., Peng, S., Kukelova, Z., Pollefeys, M. and Sattler, T.: WildGaussians: 3D Gaussian Splatting in the Wild, *NeurIPS* (2024).

[6] Li, D., Jiang, K., Tang, Y., Ramamoorthi, R., Chellappa, R. and Peng, C.: MS-GS: Multi-Appearance Sparse-View 3D Gaussian Splatting in the Wild, *NeurIPS* (2025).

[7] Lin, H., Chen, S., Liew, J. H., Chen, D. Y., Li, Z., Shi, G., Feng, J. and Kang, B.: Depth Anything 3: Recovering the visual space from any views, *arXiv preprint arXiv:2511.10647* (2025).

[8] Martin-Brualla, R., Radwan, N., Sajjadi, M. S. M., Barron, J. T., Dosovitskiy, A. and Duckworth, D.: NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections, *CVPR*, pp. 7210–7219 (2021).

[9] 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).

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

[11] Rudnev, V., Elgharib, M., Smith, W., Liu, L., Golyanik, V. and Theobalt, C.: NeRF for Outdoor Scene Relighting, *ECCV* (2022).

[12] Tung, J., Chou, G., Cai, R., Yang, G., Zhang, K., Wetzstein, G., Hariharan, B. and Snavely, N.: MegaScenes: Scene-Level View Synthesis at Scale, *ECCV* (2024).

[13] Xu, J., Gao, S. and Shan, Y.: FreeSplatter: Pose-free Gaussian Splatting for Sparse-view 3D Reconstruction, *ICCV*, pp. 25442–25452 (2025).

[14] 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).

[15] 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).

[16] Zhang, Q., Huang, C., Zhang, Q., Li, N. and Feng, W.: SU-RGS: Relightable 3D Gaussian Splatting from Sparse Views under Unconstrained Illuminations, *ICCV*, pp. 26859–26868 (2025).

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

[18] Zhang, S., Wang, J., Xu, Y., Xue, N., Rupprecht, C., Zhou, X., Shen, Y. and Wetzstein, G.: FLARE: Feed-forward Geometry, Appearance and Camera Estimation from Uncalibrated Sparse Views, *CVPR*, pp. 21936–21947 (2025).