7ZF-06

カメラポーズが未知の環境下での少ない画像からの深度画像を用いたNeRF

佐藤 和仁 † 武田 司 † 岩瀬 翔平 † 山口 周悟 † 森島 繁生 ‡
† 早稲田大学 ‡ 早稲田大学理工学術院総合研究所

1. はじめに

自由視点映像はスポーツやエンターテイメントで需要が高まっている。一般的に自由視点映像を生成するには大規模な撮影システムを必要であるため、撮影の手間とコストの軽減が求められている。

複数枚の画像から現実世界のシーンをフォトリアルに新規ビュー合成することができる Neural Radiance Fields (NeRF) [1] は注目を集めている。しかし、NeRF の制約として、(1) 視点の異なる多くの画像と (2) 正確なカメラポーズを必要とする。既存研究では、少ない画像から新規ビュー合成する手法 [2] と不正確なカメラポーズから新規ビュー合成する手法 [3] がそれぞれ提案されているが、2 つの制約を同時に解決する手法が提案されていない。

本稿では、少ない画像と不正確なカメラポーズから深度画像を用いて新規ビュー合成する手法を提案する。提案手法では、推定された深度が正解の深度に近づくように損失を追加することで、少ない画像からニューラルシーン表現とカメラポーズを同時に最適化することを目指す。

2. 関連研究

2.1 Neural Radiance Fields (NeRF)

Mildenhall ら [1] は複雑な静的シーンにおいてカメラポーズが既知の複数枚の画像から新規ビュー合成する NeRF を提案した。NeRF は MLP を使って連続的な Radiance Field を表現し、ボリュメトリックレンダリングで画像を合成する。

2.2 Bundle-Adjusting Neural Radiance Fields (BARF)

Lin ら [3] は不正確なカメラポーズから NeRF の学習を行う Bundle-Adjusting Neural Radiance Fields(BARF) を提案した。BARF は coarse-to-fine に学習を進めることでニューラルシーン表現の最適化と不正確なカメラポーズの修正を同時に行うことが可能であることを示した。

我々の手法は BARF の手法をベースとして、深度画像を用いることで少ない画像からでも BARF の学習が進められるようにした。

3. 提案手法

本稿では、少ない画像と不正確なカメラポーズから深度画像を用いて新規ビュー合成する手法を提案する。提案手法では、推定された深度が正解の深度に近づくように損失を追加することで、少ない画像からニューラルシーン表現とカメラポーズを同時に最適化することを目指す。

3.1 ボリュメトリックレンダリング

NeRF は MLPf を使って 3 次元の点 $\mathbf{x}\in\mathbb{R}^3$ と視線方向 $\mathbf{d}\in\mathbb{R}^3$ から RGB 値 $\mathbf{c}\in\mathbb{R}^3$ と密度 $\sigma\in\mathbb{R}$ を予測する。$\mathbf{\Theta}$ を MLPf のパラメータとすると、$f(\mathbf{x},\mathbf{d};\mathbf{\Theta})=(\mathbf{c},\sigma)$ とまとめられる。

RGB 画像をレンダリングするとき、6DoF でパラメータ化されたカメラポーズ $\mathbf{p}\in\mathbb{R}^6$ から各ピクセルに対応したレイ $\mathbf{r}\in\mathcal{R}(\mathbf{p})$ を生成する。t をカメラポーズ $\mathbf{p}$ の投影中心 $\mathbf{o}$ とレイ方向 $\mathbf{d}$ を使って、$\mathbf{r}=\mathbf{o}+t\mathbf{d}$ とパラメータ化すると、RGB 画像の各ピクセルの色は以下の式で推定する。

$$\hat{\mathcal{I}}(\mathbf{r})=\int_{t_{near}}^{t_{far}} T(t)\sigma(t)\mathbf{c}(t)dt \tag{1}$$

このとき、$T(t)=\exp(-\int_{t_{near}}^{t}\sigma(s)ds)$ と表され、t_{near} と t_{far} はレイ上でボリュメトリックレンダリングする区間の手前と奥の境界である。

深度画像についても、RGB 画像と同様に以下の式でボリュメトリックレンダリングで推定する。

$$\hat{\mathcal{D}}(\mathbf{r})=\int_{t_{near}}^{t_{far}} T(t)\sigma(t)tdt \tag{2}$$

3.2 損失と最適化パラメータ

与えられた M 枚の正解の RGB 画像を $\{\mathcal{I}_i\}_{i=1}^{M}$、正解の深度画像を $\{\mathcal{D}_i\}_{i=1}^{M}$ とする。提案手法では、NeRF のパラメータ $\mathbf{\Theta}$ とカメラポーズ $\{\mathbf{p}_i\}_{i=1}^{M}$ を最適化することを目的とする。損失と最適化するパラメータは以下のようにまとめられる。

$$\min_{\mathbf{p}_1,\dots,\mathbf{p}_M,\mathbf{\Theta}} \sum_{i=1}^{M}\sum_{\mathbf{r}\in\mathcal{R}(\mathbf{p}_i)} ||\hat{\mathcal{I}}(\mathbf{r};\mathbf{\Theta})-\mathcal{I}_i(\mathbf{r})||_2^2 + \lambda||\hat{\mathcal{D}}(\mathbf{r};\mathbf{\Theta})-\mathcal{D}_i(\mathbf{r})||_2^2 \tag{3}$$

3.3 Positional Encoding

Positional Encoding は NeRF がフォトリアルに合成することを可能にしている。Positional Encoding は入力された 3 次元座標 $\mathbf{x}$ を正弦波周波数ベースの高次元にマッピングする。L 個の周波数帯を持つ Positional Encoding は以下のように定義される。

$$\gamma(\mathbf{x})=(\mathbf{x},\gamma_0(\mathbf{x}),\gamma_1(\mathbf{x}),\dots,\gamma_{L-1}(\mathbf{x}))\in\mathbb{R}^{3+6L} \tag{4}$$

このとき、k 番目のエンコーディング $\gamma_k(\mathbf{x})$ 以下のように表される。

$$\gamma_k(\mathbf{x})=(\cos(2^k\pi\mathbf{x}),\sin(2^k\pi\mathbf{x}))\in\mathbb{R}^6 \tag{5}$$

最終的な NeRF のネットワークは Positional Encoding との合成 $f(\mathbf{x})=f'\circ\gamma(\mathbf{x})$ で表される。

NeRF using depth images from sparse inputs and unknown camera poses:
Kazuhito Sato†, Tsukasa Takeda†, Shohei Iwase†, Shugo Yamaguchi†, and Shigeo Morishima‡ (†Waseda University, ‡Waseda Research Institute for Science and Engineering)

提案手法では、BARF [3] と同様に coarse-to-fine の手法を導入する。学習プロセスの間、動的なローパスフィルタのように機能するマスクを Positional Encoding に適用する。k 番目のエンコーディングは以下のように表される。

$$\gamma_k(\mathbf{x};\alpha) = w_k(\alpha)\cdot(\cos(2^k\pi\mathbf{x}),\sin(2^k\pi\mathbf{x})) \quad (6)$$

$$w(\alpha) = \begin{cases} 0 & (\alpha < k) \\ \dfrac{1-\cos((\alpha-k)\pi)}{2} & (0 \le \alpha - k < 1) \\ 1 & (\alpha - k \ge 1) \end{cases} \quad (7)$$

このとき、$\alpha \in [0, L]$ は学習プロセスの間、動的に制御するパラメータである。$\mathbf{x}$ のみの $\alpha = 0$ から始めて、徐々に周波数帯を増やしていき、最終的に $\alpha = L$ とし、すべての周波数帯を活性化させる。

4. 実験

4.1 実験設定

不完全なカメラポーズを再現するために、正確なカメラポーズに正規分布 $\mathcal{N}(0, 0.15)$ に従うノイズを追加した。解像度が 400×400 の画像を用いて実験を行った。各学習ステップでは、ランダムに画像を 1 枚選択し、その画像の中から 1024 ピクセルをランダムにサンプリングし、レンダリングを行った。200000 イテレーションまで学習を行い、ネットワーク f の学習率は 1×10^{-4} から 1×10^{-5} まで、カメラポーズ $\mathbf{p}$ の学習率は 1×10^{-3} から 5×10^{-5} まで指数関数的に減衰させた。Positional Encoding に関して、20000 イテレーションから 100000 イテレーションまでの間、$\alpha = 0$ から $\alpha = L$ まで線形に変化させた。

4.2 結果

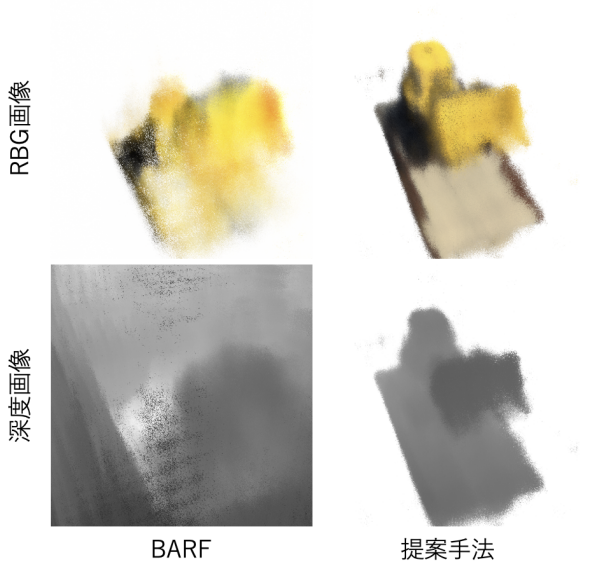

図 1: レンダリング結果の定性的比較

図 1 は正解の RGB 画像と深度画像それぞれ 4 枚を学習させたときの BARF [3] と提案手法でのレンダリングした結果である。図 2 は BARF [3] と提案手法での最適化されたカメラポーズの比較である。RGB 画像とカメラポーズの両方において、深度画像を用いない BARF の手法よりも深度画像を用いた提案手法の方が推定精度が高くなっていると思われる。

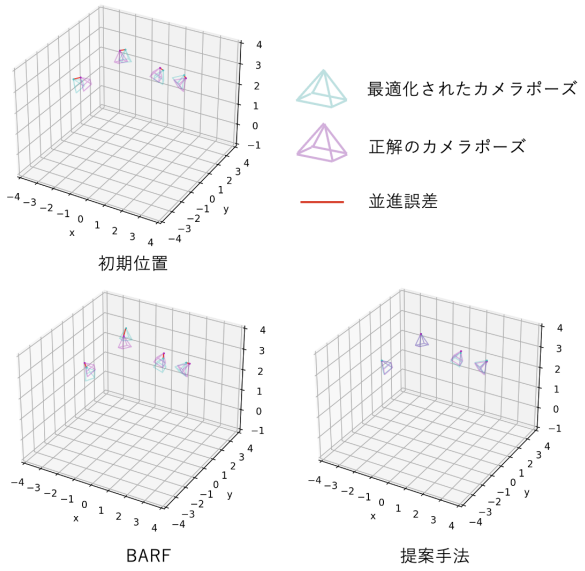

図 2: カメラポーズの定性的比較

5. おわりに

本稿では、少ない画像と不正確なカメラポーズから深度画像を用いて新規ビュー合成する手法を提案した。実験により、画像枚数が少ない場合、深度画像を用いた方がレンダリングの品質とカメラポーズの推定精度が高くなることがわかった。今後は、スマートフォンを使って現実世界のシーンでの実験を行いたい。

謝辞

この研究は、JST 未来社会創造事業 (JPMJMI19B2) および JSPS 科研費 (19H01129, 19H04137, 21H05054) の補助を受けた。

参考文献

[1] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

[2] Julian Chibane, Aayush Bansal, Verica Lazova, and Gerard Pons-Moll. Stereo radiance fields (srf): Learning view synthesis from sparse views of novel scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2021.

[3] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In *IEEE International Conference on Computer Vision (ICCV)*, 2021.