2U-07

カメラ間で時刻同期していない動画を用いた Dynamic NeRF の検討

佐々木 馨 † 佐藤 和仁 † 山口 周悟 † 武田 司 † 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

近年，撮影を行ったカメラの視点とは異なる視点からの映像を生成する，自由視点映像生成の分野が注目を集めている．この自由視点映像を生成する手法として，Neural Radiance Fields（NeRF）[5] をベースとして，時刻情報を組み込んだ Dynamic NeRF の研究も積極的に取り組まれている．Dynamic NeRF は，空間的のみならず，撮影されたフレーム間の新規時間の画像を生成することも可能である．

多視点から撮影した動画を入力とする Dynamic NeRF は，撮影時にカメラ間でタイムコード同期などの時刻同期を行い撮影された動画を用いることを前提としている．しかし，SNS 上の動画や複数人により気軽に撮影された動画は，撮影者が異なるため，カメラ間で時刻の同期がされていない非同期の動画となる．

そこで本研究では，Dynamic NeRF の入力に，オフセット（時刻ずれ）推定用の学習可能パラメータを導入することで，視点ごとのオフセットを学習する．これにより非同期動画を用いた Dynamic NeRF による自由視点映像の生成を目指す．

2. 関連研究

2.1 Neural Radiance Fields（NeRF）

Mildenhall らは，多視点から撮影した画像とカメラポーズから，静的シーンの自由視点画像を生成する NeRF [5] を提案した．NeRF は、三次元空間上の各点の位置と視線方向を入力として，RGB 値と密度を出力する輝度場を学習し，微分可能なボリュームレンダリングによって画像を合成する．

2.2 Dynamic NeRF

静的シーンを対象とする NeRF を動的シーンへ拡張した様々な Dynamic NeRF が提案されている．Li らは，輝度場の学習の入力に時刻情報を加え，さらにシーン中の速度場を学習させることにより，自由視点でフレーム間の補間画像を合成可能な NSFF [3] を提案した．また，Li らは，多視点の動画から効率的な自由視点映像を合成する DyNeRF [2] を提案した．さらに，学習の高速化に着目し，Fridovich-Keil らは，シーンをグリッドベースで学習する K-Planes [1] を提案した．これら Dynamic NeRF では，動画内全フレームに正確な時刻情報を与える必要がある．このため，多視点の動画を用いる場合，時刻同期された動画を用いることを前提としている．しかし，SNS 上の動画や気軽に撮影した動画は時刻非同期な動画であり，Dynamic NeRF を用いて高品質な自由視点映像の合成を行うことは困難である．

2.3 NeRF におけるパラメータ最適化

NeRF は，正確なカメラポーズの利用を前提としているが，撮影環境に依っては正確なカメラポーズが入手できない場合もある．そこで Lin らは，NeRF の入力であるカメラポーズ $\mathbf{p}$ を学習可能パラメータとし，輝度場を表現する MLP のパラメータ $\boldsymbol{\Theta}$ と同時にカメラポーズを最適化するアプローチを提案した [4]．M 枚の画像を用いるとき，各ピクセルの RGB 値 $\hat{\mathbf{C}}(\mathbf{r};\boldsymbol{\Theta})$ と正解の RGB 値 $\mathbf{C}(\mathbf{r})$ から式 (1) に示す損失関数で最適化を行う．

$$\min_{\mathbf{p}_1,\ldots,\mathbf{p}_M,\boldsymbol{\Theta}} \sum_{i\in M} \sum_{\mathbf{r}\in\mathcal{R}(\mathbf{p}_i)} \|\hat{\mathbf{C}}(\mathbf{r};\mathbf{p}_i,\boldsymbol{\Theta}) - \mathbf{C}(\mathbf{r})\|_2^2 \quad (1)$$

このとき，$\mathcal{R}$ はレンダリングを行う全ピクセル，$\mathbf{r}$ は各ピクセルに対応するレイを表す．この手法により，不正確なカメラポーズを入力としても，カメラポーズを修正しながらシーンを再構成することが可能である．

3. 提案手法

多視点の動画を入力として，動的シーンの輝度場を学習する際，視点ごとのオフセットを推定しながら学習を行う手法を提案する．Dynamic NeRF の入力である各フレームの時刻情報に視点ごとに学習可能なオフセット推定用のパラメータを加えることで不正確な時刻情報を修正しながらシーンを再構成することを目指す．

3.1 オフセット推定用のパラメータの追加

Dynamic NeRF は，三次元空間上の位置 $\mathbf{x}\in\mathbb{R}^3$ と視線方向 $\mathbf{d}\in\mathbb{R}^3$，各フレームの時刻情報 $t\in\mathbb{R}$ から色 $\mathbf{c}\in\mathbb{R}^3$ と密度 $\sigma\in\mathbb{R}$ を推定する MLP F_Θ のパラメータ $\boldsymbol{\Theta}$ を学習する．

$$F_\Theta : (\mathbf{x},\mathbf{d},t) \longrightarrow (\mathbf{c},\sigma) \quad (2)$$

ここで，時刻に視点ごとにオフセット推定用の学習可能なパラメータ δ_k を加える．

$$F_\Theta : (\mathbf{x},\mathbf{d},t+\delta_k) \longrightarrow (\mathbf{c},\sigma) \quad (3)$$

レンダリングを行う際，全てのピクセル $\mathcal{R}$ から各ピクセルに対応するレイ $\mathbf{r}$ を生成する．各三次元上の点は視点原点 $\mathbf{o}$ と視線方向 $\mathbf{d}$ を用いたレイ $\mathbf{r}$ を用いて $\mathbf{r}(s)=\mathbf{o}+s\mathbf{d}$ と表せられ，各ピクセルの RGB 値 $\mathbf{C}(\mathbf{r})$ は式 4 で計算する．

$$\mathbf{C}(\mathbf{r}) = \int_{s_n}^{s_f} T(s)\sigma(\mathbf{r}(s),t+\delta_k)\mathbf{c}(\mathbf{r}(s),\mathbf{d},t+\delta_k) \quad (4)$$

このとき，s_n，s_f はレンダリングするレイ上の範囲を指し，$T(s)=\exp(-\int_{s_n}^{s}\sigma(\mathbf{r}(l),t+\delta_k)dl)$ と表される．

3.2 パラメータの最適化

レンダリングされた各ピクセルの RGB 値 $\hat{\mathbf{C}}(\mathbf{r};\boldsymbol{\Theta})$ と正解の RGB 値 $\mathbf{C}(\mathbf{r})$ で二乗和誤差を計算すること

Consideration of Dynamic NeRF using videos that are not time-synchronized between cameras:
Kaoru Sasaki†, Kazuhito Sato†, Shugo Yamaguchi†, Tsukasa Takeda†, and Shigeo Morishima‡ (†Waseda University, ‡Waseda Research Institute for Science and Engineering)

で MLP のパラメータ $\mathbf{\Theta}$ とオフセット推定パラメータ δ_k を同時に学習する．以上より，各視点 $k \in \mathcal{K}$，各時刻 $t \in \mathcal{T}$，各レイ $\mathbf{r} \in \mathcal{R}$ を通して式 5 に示す損失関数で学習を行う．

$$\min_{\delta_1,\ldots,\delta_k,\mathbf{\Theta}} \sum_{k\in\mathcal{K}} \sum_{t\in\mathcal{T}} \sum_{r\in\mathcal{R}} \|\hat{\mathbf{C}}(\mathbf{r},t;\delta_k,\mathbf{\Theta}) - \mathbf{C}(\mathbf{r},t)\|_2^2 \quad (5)$$

4. 実験

4.1 実験概要

K-planes をベースにしてオフセット推定を実装し，30fps で 18 視点の動画データセット (解像度:520×520) を Blender を用いて作成した．4.1.1 では，オフセット推定を行うことで時刻非同期の動画を使用しても，高品質な画像を生成できることを示す．4.1.2 では時刻非同期の動画を用いてオフセット推定を行うことで，時刻同期動画を使用するよりも，フレーム間の補間画像の品質が向上することを示す．

4.1.1 オフセット推定

3 視点ずつオフセットを $n/270$ 秒 $(n=0,1,3,5,6,7)$ もたせ，時刻非同期動画を模倣した．各フレームの時刻情報は各視点でのフレーム番号を用い，オフセットを無視したデータセットを作成した．このデータセットに対して，提案手法によりオフセット推定を行う場合と，行わない場合について学習を行い合成画像の品質の比較を行った．

4.1.2 フレーム間の補間画像生成

3 視点ずつオフセットを $n/270$ 秒 $(n=0,1,3,5,6,7)$ もたせた時刻非同期動画に相当するデータセットと 18 視点全てにオフセットがない時刻同期動画に相当するデータセットを用いてそれぞれ学習を行った場合のフレーム間の補間画像の品質の比較を行った．

4.2 結果

4.2.1 オフセット推定

図 1 に合成画像を示す．オフセット推定を行うことで，実際には異なる時刻情報のフレームが同時刻情報のフレームとして扱われず，各フレームに対して正しい時刻情報を学習しながらシーンを再構成できることで動きのある箇所を高品質に再構成可能となった．

また，各視点において正しいオフセット値と推定されたオフセット値の平均絶対誤差は 30fps における 0.016 フレームであった．

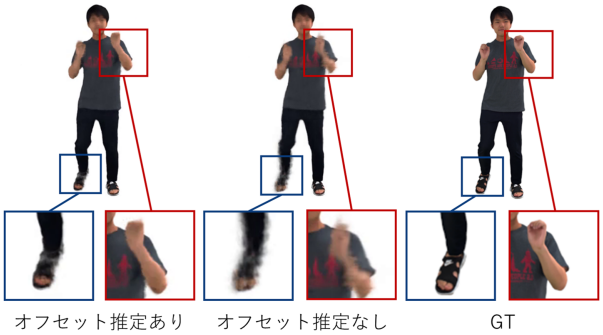

図 1: オフセット推定の有無による品質比較

4.2.2 フレーム間の補間画像生成

図 2 に合成画像を示す．多視点の動画を用いてフレーム間の補間画像を生成する際，時刻非同期動画を用いることで時刻方向の情報量が増加し，これにより対象の形状の崩壊を大幅に抑制できることが明らかとなった．

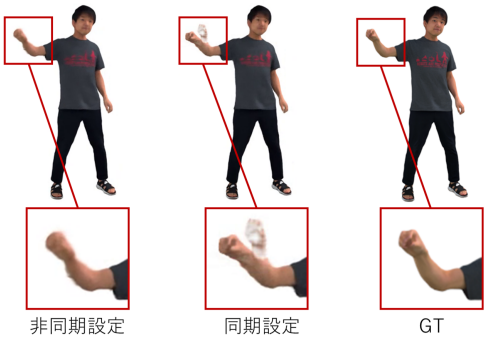

図 2: 新規時間の合成画像の品質比較

5. おわりに

本稿では，多視点の動画を入力として，動的シーンの輝度場を学習する際，視点ごとのオフセットを推定しながら学習を行う手法を提案した．実験により，提案手法によりオフセットの学習を行うことが可能であることが分かった．また，時刻非同期動画を用い，オフセット推定を行うことで，時刻同期動画を用いた場合より，フレーム間の補間画像の品質が向上することが分かった．これにより，Dynamic NeRF を用いてフレームレートを上げた自由視点映像を合成する場合，非同期動画を用いることが有効であると分かった．今後はさらに in the wild な問題設定となるカメラポーズの最適化とオフセット推定を同時に行う実験を行いたい．

謝辞

本研究は，JSPS 科研費 (21H05054) の補助を受けています．

参考文献

[1] Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. In *CVPR*, 2023.

[2] Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, and Zhaoyang Lv. Neural 3d video synthesis, 2021.

[3] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *CVPR*, 2021.

[4] 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.

[5] 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.