H-034

高頻度デジタルツインに向けた自由視点合成手法を用いた空間三次元再構築
3D Reconstruction Using Novel View Synthesis Method for High-Frequency Digital Twinning

野畑 洸貴† 石黒 祥生†‡ 大谷 健登† 西野 隆典§ 武田 一哉†‡
Koki Nobata Yoshio Ishiguro Kento Ohtani Takanori Nishino Kazuya Takeda

1. はじめに

本研究では，機械学習を利用し，複数の画像データセットからデータが含まれない新たな視点からの画像合成手法である「Neural Radiance Fields」(以下 NeRF)を Light Detection And Rangings センサー(以下 LiDAR)の位置姿勢推定を併用して学習する手法を提案する．

VR を用いたバーチャル観光や，現実のコピーを仮想空間上に作成し現実では実現の難しいシミュレーションを行うことができるデジタルツインといったサービスの需要が高まっている．これらのサービス，アプリケーションを実現する上で，現実世界を高精細に再現した仮想三次元空間再構築は重要な課題である．現実環境は常に変化する．そのため，その変化に対応するためには，高頻度なデータ収集が必要である．しかし，測量用レーザスキャナでは，高精度，高密度な点群データを収集可能である一方，定点での計測が必要であり，計測に時間がかかるため道路などでは高頻度なデータの収集は難しい．また，車載搭載の LiDAR は，走行しながらデータの収集が可能だが，収集できるのは測量用と比べ密度が疎な点群データである．

そこで，走行しながらの撮影が容易な画像と車載 LiDAR で収集できる密度の低い点群データを組み合わせる手法を提案する．それにより高精度，高密度な点群データを，高頻度に更新を可能とすることを目指す．

2. 関連研究

2.1 NeRF

NeRF[1]は，複数の視点から撮影された画像データセットと画像から推定される位置や姿勢といった視点情報から，機械学習によりそれらの場面に特化したモデルを構築し，画像データセットにない新たな視点からの画像を生成する手法である．NeRF では，Radiance Fields という，空間上のある座標の物体の密度(そこに存在しているかどうかの指標)と色を対応付けるベクトル場をニューラルネットワークで表現し，任意視点からの画像を合成する．学習には画像だけでなく，そのカメラ位置姿勢情報も必要である．視点の座標(x,y,z)と角度(θ,ϕ)を入力とし，視点情報に依らない体積密度と視点依存の色の情報により，視点に近い座標を優先して画像を生成する．視点情報は多視点画像からの 3D 形状復元技術である Structure from Motion[2]技術を利用した COLMAP[3]によって推測された位置，姿勢情報を用いている．

†名古屋大学 Nagoya University
‡株式会社ティアフォー TIER IV, Inc.
§名城大学 Meijo University

2.2 NeRF 派生研究

NeRF の原理を利用し，立体モデルに物体のスライス画像の情報を加えたモデルを構築するというレベルセット法のスライス画像を曲げることで，時間変化による対象の形状，位置の変化に対応した HyperNeRF[4]や， RGBD カメラを利用し，自己位置推定と三次元地図作成を同時に行う SLAM(Simultaneous Localization and Mapping)である iMAP[5]といった NeRF の三次元空間への利用，NeRF 自体の性能の向上などの様々な観点で次々と派生研究が発表されている．これらの研究は小規模でオブジェクト中心の再構成に焦点が当てられているが，Block-NeRF[6]では，都市を複数の Block に分割し各 Block を個別に学習して，合成時に組み合わせることで対象を都市規模にスケールアップすることにより，大規模環境においても NeRF が有効であることが示されている．

しかし，車両搭載カメラは地面に対して水平に移動し，垂直方向の位置差がなく視点位置姿勢推定の精度が落ちる可能性や，特徴点が少ない場合や画像枚数が不十分な場合には，視点位置姿勢推定自体が困難となる可能性がある．

そこで，本研究では，画像による位置姿勢推定ではなく，自動運転研究において車両自己位置推定で利用されている，LiDAR と三次元地図の形状のマッチングによる視点情報推定手法を利用する．この位置情報を利用し，HyperNeRF の学習によって得られたモデルから合成された画像の再現率が，従来の画像ベースの手法と比べて頑健かどうか検証する．画像ベースではなく LiDAR によるマッチングで推定された視点情報を利用することで，二つの利点がある．一つ目は，スケールである．COLMAP から推定される位置情報が相対位置であるのに対し，LiDAR 情報と三次元地図情報のマッチングによって得られる姿勢情報は実スケールである．二つ目は，視点位置精度が特徴点に依存しないことである．平坦な壁やトンネルなど，画像特徴量が少なく視点位置推定が困難な場面でも視点位置を取得できる．

3. 電動車椅子で収集したデータによる学習

同一画像データセットから COLMAP による視点位置姿勢推定結果(以下 COLMAP 視点情報)と，LiDAR と三次元地図とのマッチングによる位置姿勢推定(以下 LiDAR 視点情報)の二つの視点位置姿勢情報を，HyperNeRF の視点位置姿勢情報として学習を行った．

そして，各視点位置姿勢情報の精度，また，学習されたモデルからレンダリングされた画像と学習に用いた元画像との再現度の評価を行なった．そのフローチャートを図 1 に示す．

3.1 実験方法

電動車椅子 WHILL の上部に LiDAR，進行方向に対して右に 90 度回転させたカメラを設置し，画像サイズが 640×

480 の画像とその撮影時の位置姿勢情報を記録した．WHILL の自己位置情報は，自動運転ソフトウェアの Autoware[7]の自己位置推定機能を利用した．これは，LiDAR と三次元地図情報のマッチングにより自己位置を推定する．学内の一般的なオフィス環境を，6 ケースに分けて走行した．1ケースあたりの移動距離はおよそ5から10mである．4 ケースを直線，2ケースを曲線で走行した．壁面を対象として 4 ケース，部屋の内部を対象として 2 ケース走行した．

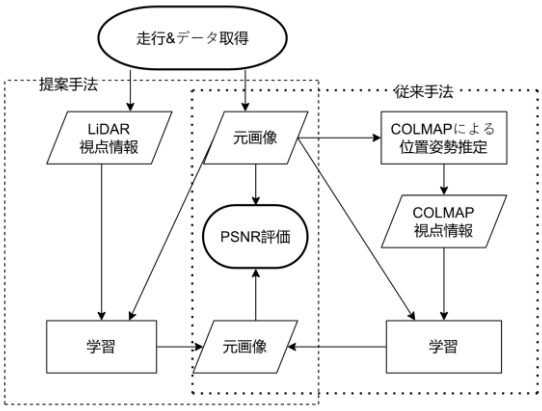


図 1　再現度評価のフローチャート

3.2 評価方法

LiDAR 位置姿勢推定情報を用いた NeRF により再構成した画像の PSNR を評価データとして用いる．比較として，従来手法の COLMAP 位置姿勢推定も行い，LiDAR 位置姿勢推定の場合との違いを画像再構成時の PSNR(Peak-Signalto Noise Ratio)を用いて評価する．

移動が直線の走行において，各視点位置姿勢情報の位置が完全なる直線であると仮定し，回帰直線との距離の MSE(Mean Squared Error)で視点位置の精度を評価する．MSE とは，予測値と正解値の差の二乗値であり，MSE が小さいほど位置推定精度が良い．また，姿勢においても変化がないと仮定し，姿勢情報の分散で姿勢について評価する．分散は低いほど姿勢推定精度は高い．

3.3 各視点位置姿勢精度

COLMAP，LiDAR を用いて HyperNeRF に用いる視点情報を取得した．直線の走行は，位置座標が直線で姿勢は変化していないと仮定し，各視点情報取得手法での直線走行ケースにおける各視点位置姿勢情報の画像枚数，推定位置とその回帰直線との MSE，推定姿勢の分散を表 1 に示す．

COLMAP 視点情報の場合の方が LiDAR 視点情報より，直線上を走行した場合の位置の MSE と姿勢の分散がほとんどの場合で小さく，位置姿勢推定が高い結果となった．

表 1　走行ケースと各位置情報の MSE，各姿勢情報の分散

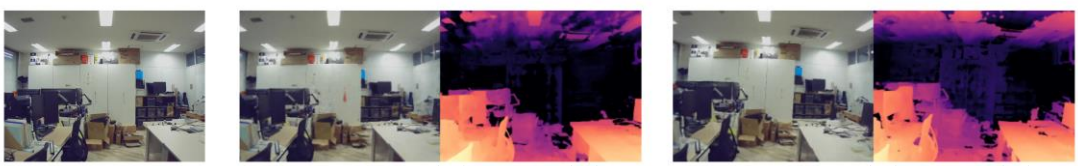
走行	位置の MSE		姿勢の分散	
	COLMAP	LiDAR	COLMAP	LiDAR
1	1.61e-03	1.29e-02	3.79e-04	5.94e-04
2	1.13e-04	1.32e-01	4.79e-05	1.11e-04
3	曲線			
4	9.91e-04	6.57e-02	2.31e-04	2.13e-04
5	曲線			
6	7.93e-05	7.79e-02	3.70e-05	9.76e-04

3.4 実験結果

1 ケース分のデータセットでの学習時間はおよそ 8 時間であった．

各走行の入力画像とそれぞれの視点情報で学習したモデルから，合成された RGB，Depth 画像を図 2 に示す．天井やロッカーといった平面の部分で，色の再現はできているが，Depth 画像では穴があるような結果であった．

表 2 に各視点位置姿勢情報における学習結果からの合成画像と元画像との PSNR を示す．平均，最大値，最小値では全ての走行ケース，学習に用いる視点情報が COLMAP 視点情報の場合が LiDAR 視点情報よりも再現度が高い結果となった．平均や最小値では，撮影対象が壁面である走行ケース 1,2,4,5 において LiDAR 視点情報を用いた場合の PSNR が COLMAP 視点情報の 80%以下となる走行が多い．最大値では，同様に COLMAP 視点情報の場合は PSNR が高いが，平均や最小値ほど各視点情報による差が少ない．しかし，分散は LiDAR 視点情報の方が良い場合が多い．また，走行ケースごとの PSNR の変化量は LiDAR 視点情報の場合の方が安定している．



入力画像　COLMAP:合成 RGB(左)，Depth(右)画像　LiDAR:合成 RGB(左)，Depth(右)画像

図 2　入力画像と各視点情報によるモデルからの学習

表 2　走行ケースと各視点情報を用いた合成画像の PSNR

走行	画像枚数	COLMAP				LiDAR(COLMAP 視点情報との増減率)			
		平均値	最大値	最小値	分散	平均値	最大値	最小値	分散
1	91	26.4	32.6	22.4	7.03	15.3(-42.1%)	29.3(-10.1%)	11.9(-46.9%)	7.41(+5.4%)
2	94	23.8	32.6	19.7	6.83	19.0(-20.2%)	31.0(-4.91%)	16.1(-18.3%)	4.45(-34.8%)
3	84	18.6	30.4	16.0	9.92	14.0(-24.7%)	24.4(-19.7%)	11.3(-39.2%)	8.96(-9.68%
4	92	18.8	27.4	17.0	5.58	16.6(-11.7%)	26.0(-5.10%)	10.7(-37.1%)	3.19(-42.9%)
5	84	20.8	25.3	19.6	0.957	12.8(-38.5%)	24.2(-4.34%)	10.7(-45.4%)	5.07(+430%)
6	80	16.9	25.9	13.4	9.34	13.9(-17.8%)	25.6(-1.16%)	11.9(-11.2%)	3.55(-62.0%)

入力画像　　合成 RGB 画像　　合成 Depth 画像

図 3　車載カメラ入力画像とそれを入力データとしたモデルからの合成画像

3.5 考察

表 2 より，COLMAP 視点情報よりも，LiDAR 視点情報による学習結果の方が合成画像の再現度が低い結果となった．LiDAR による位置姿勢推定では，位置情報の更新が 10Hz であり，撮影と完全に同期しているわけではない．また，表 1 より，LiDAR とカメラの位置関係のキャリブレーションが甘かった可能性もある．COLMAP を用いた画像による位置推定では，それぞれの撮影時の位置を推定でき，カメラと他のセンサ間のキャリブレーションの必要もない．このため，LiDAR による場合も，位置情報の推定精度やキャリブレーションの精度を向上させることで結果が改善する可能性は十分考えられる．

また，対象とカメラの距離が 5m 以上離れている場合は COLMAP の場合，精度の低下が目立った．LiDAR による手法は，広い空間や周囲に画像特徴点が少ない環境でも安定して位置姿勢が推定でき，その位置情報も実スケールであるが，COLMAP 視点情報は画像データセットから推測される相対的な値であるため，実際のスケールが大きくなった場合に，誤差が大きくなるためだと考えられる．

そして，合成画像の外周部にノイズが存在する場合があり，RGB 画像では元画像と同様に合成できている場合でも Depth 画像では合成できていない場合があった．これは，画像間の垂直方向による差がなく，単方向からの視点からしか学習できていないため，その点のモデル上の密度が不安定になっていると考えられる．

4. 車載カメラで収集したデータによる学習

本実験では，自動車で収集されたデータを用いて， 第 3 章と同様に HyperNeRF の LiDAR 視点情報を用いて学習を行なった場合の合成画像の再現度の評価を行なった．

4.1 実験方法

幹線道路を走行して収集されたデータセットを利用した．車両には車体上部に LiDAR，進行方向を向いたカメラ（Center）とそれを進行方向に対して左右それぞれ 45 度方向を向いたカメラ(Left，Right)がそれぞれ設置され，画像サイズが 2880×1860 ピクセルの画像とその撮影時の位置姿勢情報が記録されている．

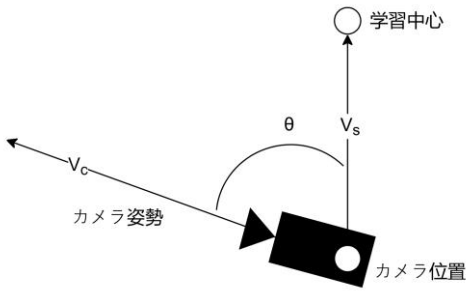


図 4　データ選択方法

4.2 入力データの選択

画像再構成したい対象の領域(建物など)を指定することにより，長時間の走行データセットの中から必要なデータを抽出する．

そこで，図 4 のように，学習の中心となる座標を決め，カメラ座標から学習の中心となる点までのベクトル Vs，カメラの姿勢ベクトル Vc を設定した．Vs の大きさが閾値以下，Vs と Vc の余弦類似度が閾値以上となるカメラ位置姿勢情報と，該当カメラ画像を複数カメラデータセットから取り出して学習の入力データとした．図 5 に LiDAR で取得された点群データの俯瞰図と学習中心を示す．

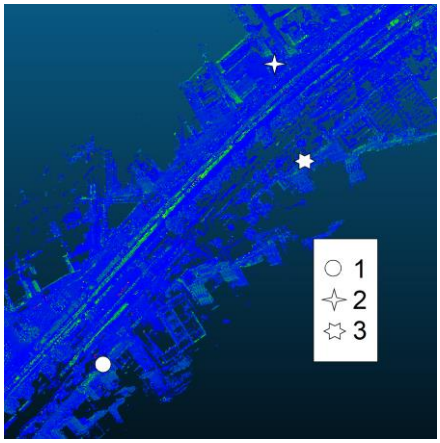


図 5　点群データと学習中心

4.3 実験結果

各学習中心における新視点からの合成 RGB 画像と元画像の PSNR を表 3 に示す．図 3 に合成された画像の一例を示す．図 2 の室内データセットの結果と比較すると，形状が室内データセットよりも不明瞭となっている．また，学習時に計算される範囲が広いため図 3 の合成 Depth 画像では明るく出力されている．

表 3　各学習中心と画像枚数，新視点合成画像の PSNR

学習中心	画像枚数	PSNR
1	21	16.4
2	31	18.9
3	28	19.3

4.4 考察

表 3 より，車載カメラデータセットで学習した結果は，室内データセットが入力の場合よりも合成画像の形状が不明瞭であるが PSNR は高い結果となった．

これは室内環境と室外環境の違いによるものだと考える．室外環境では，建物といった対象の規模が室内より大きいことと，道路といった色差が少ない対象であるという違いがある．また，他の車や歩行者などの時間によって変化する動的オブジェクトがあること，室内環境は壁で区切られているため対象までの距離が定まるが，室外では空といった無限遠の対象も学習対象としてしまうことが挙げられる．

さらに，撮影時の車両走行速度が速いためブレ画像であること，同一の撮影間隔であっても画像間の距離が大きくなることが挙げられ，データ取得方法の改善が必要である．

5. おわりに

本研究では，先行研究の HyperNeRF の学習に利用されている COLMAP 視点情報を LiDAR 視点情報に置き換え，HyperNeRF の学習を行い，元画像との再現度評価を行なった．実験結果より，LiDAR 視点情報は，COLMAP 視点情報と比べ，全体的な再現度は低い結果となったが，ケースごとの安定性の面で優位であった．よって，特徴点の少ない撮影シーンでも精度が低下することなく学習できるということが分かった．

また，LiDAR 視点情報の精度を高めることで，COLMAP での視点位置推定が困難な場合でも，安定して HyperNeRF の学習が行えることが挙げられる．車載カメラデータセットでは，走行の速度や動的オブジェクトの影響を受けており室内データセットと比べ，見かけの再現度が低い結果となった．

今後の課題としては，各学習に時間がかかったため，処理の時間を短縮することで，さらなる高頻度なデータ更新を可能とすることであるが．これについては， Instant NGP[8]などの高速化された NeRF 派生研究の利用や，学習時の入力データとして，点群データを追加することで，学習時間の効率化や時間モデルの精度を向上させることができると考える．また，本研究では，PSNR という二次元での再現度の評価を行なった．しかし，異なる環境における PSNR は合成画像から受ける印象と異なる結果となるのに加え，本研究の目的は，高密度，高精度，高頻度な三次元情報の更新である．そのため，評価手法として，複数のモデルから一つの三次元 Depth 情報の構築をし，従来の三次元取得手法で得られた三次元空間との比較が挙げられる．

謝辞

最後に，本研究を進めるにあたり株式会社マップフォーの皆様にはデータの提供，研究への助言等の多大なる支援を頂きました．心より感謝申し上げます．

参考文献

[1] Ben Mildenhall et al, NeRF: Representing scenes as neural radiance fields for view synthesis. In European conference on computer vision, (2020).
[2] J.L. Schönberger et al. Structure-from-Motion Revisited. In Conference on Computer Vision and Pattern Recognition (CVPR), (2016)
[3] J.L. Schoenberger. COLMAP Structure-fromMotion and Multi-View Stereo(Version 3.7).
[Source code]https://github.com/colmap/colmap.git. (2022).
[4] K. Park et al. HyperNeRF: A higher-dimensional representation for topologically varying neural radiance fields. arXiv preprint arXiv:2106.13228, (2021)
[5] Edgar Sucar, et al.. iMAP: Implicit mapping and positioning in real-time. arXiv preprint arXiv:2103.12352, (2021).
[6] Matthew Tancik, et al. Block-NeRF: Scalable large scene neural view synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) , (2022)
[7] Autoware.AI. Autoware-AI Open-source software for self-driving vehicles.
[Source code]https://github.com/Autoware-AI/autoware.ai.git. (2019)
[8] Thomas Müller, et al. Instant neural graphics primitives with a multiresolution hash encoding. arXiv preprint arXiv:2201.05989, (2022)