第 27 回日本バーチャルリアリティ学会大会論文集（2022 年 9 月）

自己注意機構に基づく空撮多視点画像から復元した三次元点群の欠損補完法の検討

A Study of Deficit Completion Method for 3D Point Cloud Recovered from Aerial Multi-View Images Based on Self-Attention Mechanism

木山傑将 [1)]，宍戸英彦 [2)]，北原格 [2)]
Takenobu KIYAMA, Hidehiko SHISHIDO, and Itaru KITAHARA

1) 筑波大学 理工情報生命学術院（〒305-8573 茨城県つくば市天王台 1-1-1, kiyama.takenobu@iit.tsukuba.ac.jp）
2) 筑波大学 計算科学研究センター（〒305-8577 つくば市天王台 1-1-1, {shishido | kitahara}@ccs.tsukuba.ac.jp）

概要：ドローン技術の発展に伴い空撮画像の撮影が容易になったことにより、防災・土木分野での活用や実世界コンテンツ生成等を対象にした空撮多視点画像からの三次元点群復元手法に注目が集まっている。市街地などを空撮多視点画像に Structure from Motion(SfM)を適用し三次元点群復元を行う際の問題点として、撮影位置が上空に限定されることによって発生する家屋の軒下部分の欠損や、瓦屋根などの同じような模様が繰り返されることによって発生する特徴点のマッチング誤差による欠損などが挙げられる。そこで本研究では、自己注意機構で構成した深層ニューラルネットワークにより、多視点画像から生成された点群に含まれる欠損部分を補完し整形する手法について検討する。学習フェーズでは、ドローン空撮を想定したカメラワークによって CG モデルをレンダリングした多視点画像に三次元フォトグラメトリを適用し、欠損を含む学習用三次元点群を生成する。CG モデルから生成した真値となる点群を教師データとして欠損領域を補完する深層ネットワークを学習する。復元フェーズでは、復元した三次元点群に学習ネットワークを適用することで欠損に起因する三次元復元誤差を軽減する。

キーワード：三次元点群、三次元フォトグラメトリ、深層学習、自己注意機構、欠損補完

1. はじめに

本研究では、空撮多視点画像から復元した欠損を含む三次元点群に対し、自己注意機構を活用することによって欠損箇所を補完する深層学習法を提案する。

ドローンの普及により、防災や土木開発等の分野において、空撮映像からの実世界の三次元形状復元に注目が集まっている。ドローンを遠隔操作することで三次元形状復元に必要な情報が取得可能であり、調査者が現場に足を運ぶ負担を軽減できる。さらに、Structure from Motion (SfM) [1]に代表される三次元フォトグラメトリ技術の導入によって、LiDAR 等の特殊センサーが不要となり軽量・低コスト化が可能となる。一方で、空撮多視点画像からの三次元復元には、プライバシ保護の観点から上空からの撮影が条件付けられることが多い。その結果、家屋の庇やブロック塀によって遮蔽される領域が発生し、図 1 に示すように、三次元点群が欠損するといった問題が存在する。

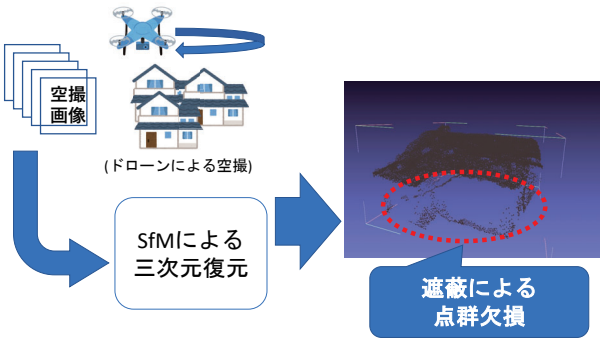


図 1: 空撮多視点画像から復元された三次元点群の欠損

本研究では、自己遮蔽によって発生する復元三次元点群の欠損に対し、自己注意機構を用いた深層学習手法を適用することにより欠損点群を補完する手法を提案する。

2. 関連研究

2.1 Structure from Motion(SfM)

SfM は、多視点画像間に存在する対応点情報から、カメ

ラ位置・姿勢の推定を行い、撮影された物体の三次元情報を復元する手法である。対応点探索には SIFT [2]等の画像間の特徴点を検出する手法を用いる。エピポーラ幾何における基礎行列からカメラ間の移動と回転を表現する行列を算出することで、カメラ位置姿勢を取得する。カメラの位置姿勢情報に基づき多眼ステレオの原理に基づき対応点の三次元情報を推定し、三次元点群を生成する。復元対象物体が自他の物体により遮蔽され撮影されていない場合には、その領域の三次元点群の復元が困難であり、点群が大きく欠損する。本研究では、そのような復元時に発生する欠損を深層学習を用いて補完する手法を提案する。

2.2 自己注意機構

自己注意機構（Self-Attention）[3]は、従来、自然言語処理分野において単語間の注意関係を示すために用いられる手法であり、他言語翻訳や文章生成など様々なタスクで用いられてきた。特に Transformer [3]は、広域・近傍を問わない参照構造を作ることで、特徴取得領域がカーネルサイズに依存する CNN（Convolution Neural Network）の課題点を解決し、画像認識や画像生成といったコンピュータビジョン分野での活用も進んでいる。

2.3 深層学習を用いた三次元点群処理

PointNet [6]は三次元点群を扱う深層学習の代表的手法であるが、点群の順不同性に対応する仕組みが複雑であること、ネットワークの入出力が固定長であることが課題であった。順不同・可変長データからの特徴取得が可能な Transformer を導入することで上述した課題を解消する Point Cloud Transformer [4]が提案されている。本研究では、Transformer を用いた点群補完手法を用いることで、SfM によって発生する点群の欠損を補完する。

本研究では、Transformer を用いた点群補完手法を用いることで、SfM によって発生する点群の欠損を補完する。

2.4 点群補完手法

点群を扱う深層学習手法には、点群をボクセル化することにより点群特徴を取得する手法[5]や、PointNet [6]のように複数の MLP(Multi-Layer Perceptron)と Maxpooling を組み合わせることにより、点群の順不同性に対応した形状特徴を抽出する手法などが存在し、それらを用いた点群補完手法が提案されている。一方で、ボクセル化による計算コストの増加や、細長い紐のような物体や複雑な形状を持つ物体など、補完対象となるオブジェクトの種類によって補完精度が低するなどの課題点が存在する。本研究では、PoinTr [7]といった自己注意機構に基づく点群補完手法を用いることで、点群の非順序性に対応した特徴抽出によって上記課題を解消した上で点群を補完する。

3. 提案手法

本研究では、空撮多視点画像に対して SfM を適応し、復元された点群に含まれる欠損を補完する深層学習手法を提案する。

3.1 点群補完を行う対象

図 1 に示すように、空撮多視点画像に SfM に適応する際には、撮影位置が上空に限定されることにより自己遮蔽が引き起こされ、復元される点群に欠損が発生する。家屋の軒下部分は、屋根に覆われ復元が困難な領域である。我々は、欠損補完を行う対象として家屋などの建造物に着目し、点群補完法を考案した。

3.2 三次元 CG モデルを用いた点群生成

点群補完を行う深層学習ネットワーク（以下、ネットワーク）の学習には、三次元復元誤差を含んでいる点群と、その真値となる正確な形状に基づく点群が必要である。実際の空撮多視点画像を用いて作成される点群の真値として、実在する建造物の正確な形状情報を取得することは困難であるため、本研究では、図 2 に示すように、建造物の三次元 CG モデルを用いて”SfM による欠損を含む点群”と”真値に該当する欠損を含まない点群”を作成する。

SfM による点群は、CG 空間中に配置した建造物の三次元モデルを上部から撮影した画像群から生成する。モデル上

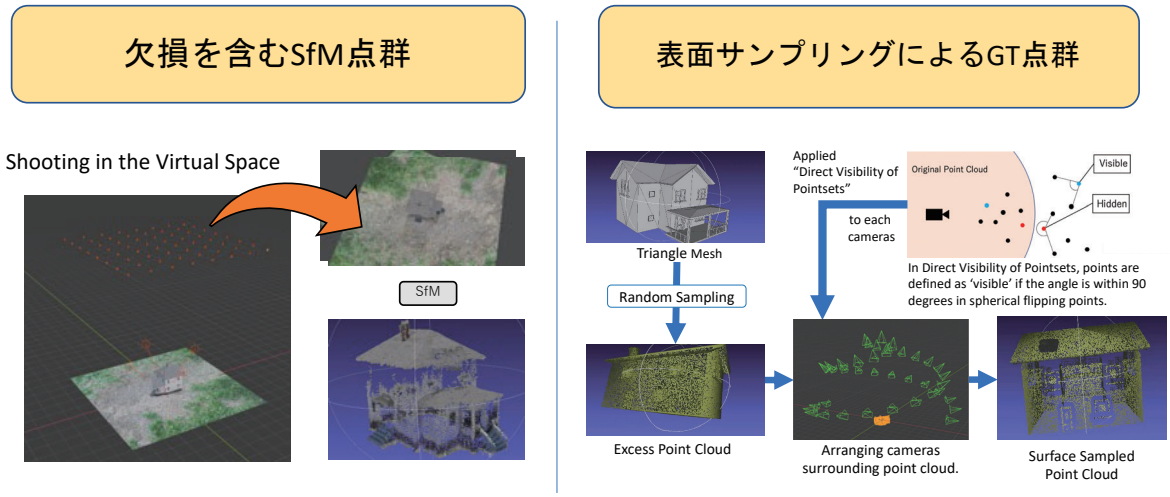


図 2 : 左) 空撮を模倣したカメラ配置により多視点画像をレンダリングし SfM を適応
右) SfM 点群の真値として 3 次元 CG モデルの表面サンプリングによって点群を生成

部に複数の仮想カメラを複数設置し、レンダリングによって得られる画像群に対して SfM を適応することで、 図２左に示すような軒下等に欠損を持つ点群を生成する。

一方、真値に該当する点群は、モデル表面のサンプリング処理によって、建造物の正確な形状情報を取得することで生成する。前述した三次元 CG モデルの表面上で、総面積に対して一様に点を打ち、三次元点群をサンプリングする。SfM で復元される点群は外表面のみであるため、建造物の内部にもメッシュを保持している場合には、そこからサンプリングされる点群を取り除く必要がある。そこで、表面サンプリングによって生成された点群の周辺に仮想カメラ配置し、Direct Visibility of Point Sets [8]を適応することで外表面上の点群のみを取得する。

4. 点群補完ネットワーク

SfM による欠損点群の補完には、Transformer のエンコーダ・デコーダを用いた点群補完手法である PoinTr を用いる。概要を図３に示す。PoinTr は欠損点群を入力とし、欠損部分を疎な点で補完したのちに、FoldingNet [9]によるアップサンプリングを行うことで整形された点群を出力するネットワークである。

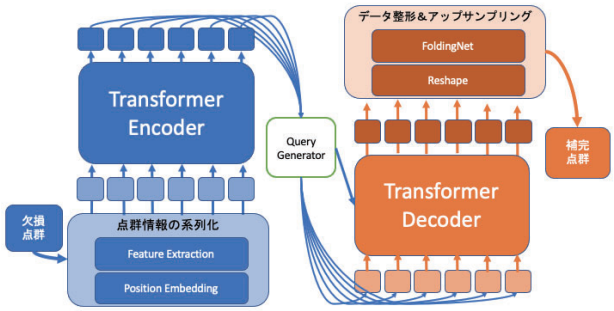


図 3: 点群補完手法 PoinTr のネットワーク概要

4.1 欠損点群の系列化

欠損を含む点群をネットワークの入出力に対応するように系列的なデータに変換する。欠損点群を入力とし、DGCNN[10]による局所特徴の取得(Feature Extraction)と、MLP による各特徴量に対する位置情報の付加(Position Embedding)を行い、点群を特徴量の系列データに変換する。その後、系列化されたデータを Transformer エンコーダに入力することで、全体特徴の取得を行う。

4.2 Query Embedding

通常の Transformer のネットワーク構造では、デコーダで用いる Query の生成を、真値となるようなデータに対し Position Embedding と注意機構を組み合わせることで実現する(Cross Attention)。

これに対し本手法では、Query Generator と呼ばれる機構を用いることで Query の生成を行う。エンコーダから出力された全体特徴から、補完する点群と同次元のベクトルを生成することで、それを Query とし、全体特徴と共に MLP に入力することで Query Embedding を行う。

4.3 点群アップサンプリング

デコーダから出力される点群は欠損部分の疎な点群であるため、密な点群生成に FoldingNet を用いた点群のアップサンプリングを用いる。FoldingNet は、点群の格子状パッチを所望の形状に変形させて、より繊細な点群表現を行う手法である。この格子状パッチを、デコーダから出力される各点を中心として貼り付け、真値となる点群に近づくように変形させることを図る。これにより、より繊細な点群表現と点群のアップサンプリングが行われる。

5. 損失関数

損失関数には、ネットワークによって出力された点群と、3.2 で述べた表面サンプリングによって生成された点群gとの Chamfer Distance を用いる。Chamfer Distance は、二つの点群間の位置誤差を図る指標で、各点から見て相手の点との最近傍距離の平均を取ることで求める。

ネットワークによって生成される点群には、デコーダの出力である欠損補完を行った疎な点群Cと、FoldingNet によりアップサンプリングされた密な点群Pの二種類がある為、式(1)のように、各々について Chamfer Distance を算出する。

$$J_0 = \frac{1}{n_{\mathcal{C}}} \sum_{c \in \mathcal{C}} \min_{g \in \mathcal{G}} \|c - g\| + \frac{1}{n_{\mathcal{G}}} \sum_{g \in \mathcal{G}} \min_{c \in \mathcal{C}} \|g - c\|$$
$$J_1 = \frac{1}{n_{\mathcal{P}}} \sum_{p \in \mathcal{P}} \min_{g \in \mathcal{G}} \|p - g\| + \frac{1}{n_{\mathcal{G}}} \sum_{g \in \mathcal{G}} \min_{p \in \mathcal{P}} \|g - p\| \quad (1)$$

J_0は疎な点群、J_1は密な点群での Chamfer Distance をそれぞれ表しており、ネットワークの最終的な損失関数としては、それらの和である$J_0 + J_1$で定義する。

6. 検証実験

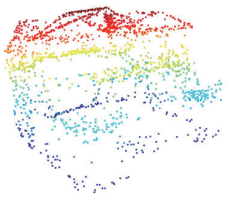
6.1 実験データ作成環境

学習データの生成に用いる建造物の三次元 CG モデルは、プラットフォームである Free 3D[11]と TurboSquid[12]から 83 種類のモデルを収集した。また、仮想空間上での空撮模倣撮影には、三次元 CG ソフトウェアである Blender[13]を使用し、収集した三次元 CG モデルを複数の仮想カメラで上部から撮影することで多視点画像の生成を行った。得られた多視点画像に対しての SfM の適応には、写真測量ソフトウェアである Pix4Dmapper[14]を使用した。真値となるような点群の生成には、Python の三次元データ処理ライブラリである Open3D[15]を用いて三次元 CG モデルの表面サンプリングを実施した。

6.2 点群補完ネットワークの学習

ネットワークの学習には、訓練用の点群データが少ないことから、三次元 CG データセットである ShapeNet[16]から作成された点群データで事前学習を行い、それを Fine Tuning する。尚、ShapeNet を用いた事前学習については、Xumin Yu ら[7]の手法を用いた。

6.1.で採集した三次元CGモデルは、訓練データ用に80種類、テストデータ用に3種類とし、それぞれから生成される点群データの点群数は、SfMによる欠損点群が2048点、真値となる表面サンプリングによる点群が8192点になるようにランダムサンプリングを行った。ネットワークの出力する点群数は、真値となる点群と同数なるように設計する。図4に、Fine Tuningにより学習を200epoch行ったテストデータの出力例を示す。



(Input : SfMによる欠損点群)

(GT : 真値となる点群)

(Output : 本手法による点群補完)

図 4: テストデータでの点群補完の出力例.

7. 考察

図4を見ると、入力されるSfMによる点群は軒下部分において、1.で提起した自己遮蔽による欠損が発生しているのに対し、本手法による出力は欠損部分に点群が補完されていることが確認できる。一方で、屋根上部や地面付近において、真値となる点群では確認できない部分での点群追加がされており、推定誤差を引き起こしている。また、全体的な形状として、明瞭な凹凸がなくなってしまっていることも確認できる。これは、事前学習を行った際に建造物以外のカテゴリ形状を学習しため、多くの種類で損失関数が少なくなるように丸みを帯びた予測形状になってしまっているのではないかと考察する。

8. おわりに

本稿では、自己遮蔽によって発生する復元三次元点群の欠損に対し、自己注意機構を用いた深層学習手法を適用することにより欠損点群を補完する手法を提案した。建造物の三次元CGモデル用いた学習データの生成により、ネットワークの訓練を行うことで、軒下などの三次元復元の際に発生する欠損について点群補完が確認できた。

一方で、本来の形状とは異なる部分に点を追加してまう出力結果や、全体的に丸みを帯びた出力結果になってしまう出力結果なども確認された。これらの課題点に対しては、学習データ数を増やすことにより、事前学習の与える影響を少なくすることで精度向上が期待される。

参考文献

[1] N.Snavely, S.M.Seitz, R.Szeliski, “Photo Tourism: Exploring Photo Collections in 3D,” ACM Transactions on Graphics, Vol.25, pp.835-846

[2] David G. Lowe, “Distinctive image features from scale-invariant keypoints,” International Journal of Computer Vision, Vol. 60, No.2, pp.91-110, 2004.

[3] Vaswani Ashish, Shazeer Noam, Parmar Niki, Uszkoreit, Jakob, Jones Llion, Gomez Aidan N., Kaiser Lukasz, Polosukhin Illia, “Attention Is All You Need”, 31st Conference on Neural Information Processing Systems (NIPS 2017)

[4] Meng-Hao Guo, Jun Xiong Cai, Zheng Ning Liu, Tai Jiang Mu, Ralph R. Martin, Shi-Min Hu, “PCT: Point cloud transformer”, Computational Visual Media, volume 7, pages187–199, 2021

[5] Charles R. Qi, HaoSu, Matthias Niessner, Angela Dai, Mengyuan Yan, Leonidas J. Guibas, “Volumetric and Multi-View CNNs for Object Classification on 3D Data”, CVPR 2016, page5648-5656

[6] Charles R. Qi, Hao Su, Kaichun Mo, Leonidas J. Guibas, “PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation”, CVPR2017 , page652-660

[7] Xumin Yu, Yongming Rao, Ziyi Wang, Zuyan Liu, Jiwen Lu, Jie Zhou, “PoinTr: Diverse Point Cloud Completion With Geometry-Aware Transformers”, International Conference on Computer Vision (ICCV), 2021, pp. 12498-12507

[8] Katz, Sagi, Tal, Ayellet, Basri, Ronen, “Direct visibility of point sets”, ACM Transactions on Graphics, Vol. 26, No. 3, Article 24, 2007

[9] Yang, Y., Feng, C., Shen, Y., Tian, D., "FoldingNet: Point Cloud Auto-encoder via Deep Grid Deformation", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018

[10] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E. Sarma, Michael M. Bronstein, Justin M. Solomon, “Dynamic Graph CNN for Learning on Point Clouds”, ACM Transactions on Graphics (TOG), 2019

[11] Free3D, https://free3d.com/(access:2022.07.13)

[12] TurboSquid, https://www.turbosquid.com/(access:2022.07.13)

[13] Blender, https://www.blender.org/(access:2022.07.13)

[14] Pix4D, https://www.pix4d.com/ (access:2022.07.13)

[15] Open3D, http://www.open3d.org/ (access:2022.07.13)