

Exploring Attention-based Deep Learning methods for Classification, Retrieval and Shape Completion of ALS Roof Point Clouds

Dimple A Shajahan*
Indian Institute of Technology Madras
TKM College of Engineering, Kerala
dimple@tkmce.ac.in

Mukund Varma T*
Indian Institute of Technology Madras
mukundvarmat@gmail.com

Ramanathan Muthuganapathy
Indian Institute of Technology Madras
mraman@iitm.ac.in

Abstract

Interpretation of Airborne Laser Scanning (ALS) point clouds, specifically building roof modeling has many applications. Existing methods do not generalize well across various roof shapes, are time-consuming, and rely on manual intervention across various stages like handcrafted feature generation, etc. In this study, we analyze deep learning-based methods specifically by incorporating attention for various roof modeling tasks including roof style classification, roof retrieval, and damaged roof completion. The proposed networks achieve state-of-the-art (SOTA) results in roof modeling tasks while still maintaining competitive performance in synthetic benchmark datasets. Our experiments indicate that attention might be an even more natural fit for point cloud processing due to its inherent permutation, cardinality invariance.

CCS Concepts: • Computing methodologies;

Keywords: attention, classification, retrieval, completion

ACM Reference Format:

Dimple A Shajahan, Mukund Varma T, and Ramanathan Muthuganapathy. 2021. Exploring Attention-based Deep Learning methods for Classification, Retrieval and Shape Completion of ALS Roof Point Clouds. In *Proceedings of ACM Conference (Conference'17)*. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/nnnnnnnn.nnnnnnnn>

1 Introduction

Representation of ALS building roof point clouds has many applications in Geographic Information Systems (GIS) [10, 20], Remote Sensing [10], Photogrammetry [10, 18] and,

Computer Vision [2, 6, 7, 10, 14]. Existing methods for these applications can be broadly classified into model-driven and data-driven methods [15]. Model-driven methods utilize a pre-defined catalog of basic roof shapes, from which the closest matching model is selected but they strongly rely on prior information about the roof style and require a vast collection of building shapes as templates to generalize well [5, 6, 12]. Data-driven methods include geometric, Machine Learning (ML), and more recently Deep Learning (DL) techniques. Geometric methods utilize basic shapes such as planes to best fit the input roof point cloud [4, 11] while ML methods process hand-crafted features through simple models like random forest classifiers, etc [1, 9]. Both these methods cannot generalize to a large set of complex roof shapes and are time-consuming [16, 17, 19]. DL has shown great performance across various computer vision tasks including deriving point cloud representations [8] and most methods focus only on analyzing clean, aligned point clouds derived from synthetic CAD models. ALS point clouds contain various imperfections like noise, outliers, missing regions, sparsity, and existing methods need not showcase strong performance in these cases [3]. Therefore, there is a strong requirement to understand the effectiveness of deep learning for ALS applications, and also propose robust models which work well in real and synthetic benchmarks.

Inspired by the recent success of attention across various Natural Language Processing tasks, our study focuses on incorporating attention to complement existing networks and also introduces full-blown attention-based networks for classification, retrieval, and shape completion of point clouds. Our key motivation is that attention is a set operator making it appropriate for processing point clouds as they are permutation and cardinality invariant. We performed detailed robustness tests to evaluate the effectiveness of the proposed methods and ensure that they are computationally efficient for real-time processing. The remainder of this report is structured as follows - Section. 2 briefly discusses the three proposed methods, Section. 3 describes key results and inferences, and finally Sections. 4, 5 concludes and suggests scope for future work.

*Both authors contributed equally to this research.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Conference'17, July 2017, Washington, DC, USA

© 2021 Association for Computing Machinery.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...\$15.00

<https://doi.org/10.1145/nnnnnnnn.nnnnnnnn>

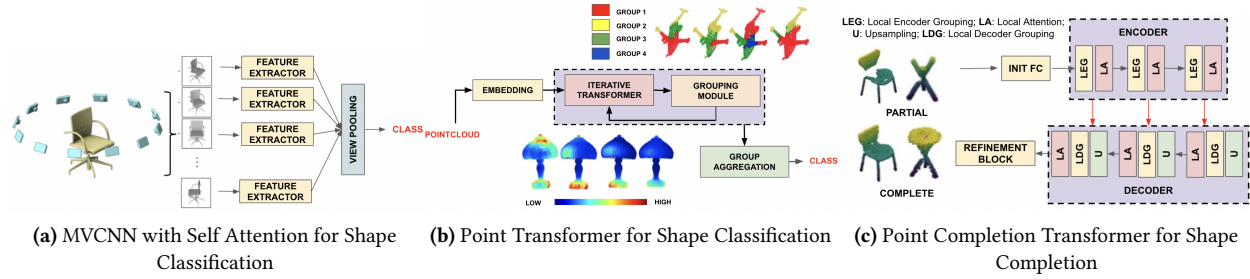


Figure 1. A brief overview of methods proposed in this report. (The figures do not show roof point clouds as their intricate geometries are difficult to visualize as a 2D image.)

2 Proposed Methods

To set some preface, we first introduce attention in subsection. 2.1 and then briefly describe the three proposed methods - Multi-view Convolutional Neural Network with Self Attention (MVCNN-SA), Point Transformer (PT), Point Completion Transformer (PCT) in the following three subsections 2.2, 2.3 and 2.4 respectively.

2.1 Attention

Attention was first proposed for NLP, where the goal is to focus on a subset of important words. Consequently, relations between inputs are highlighted that can be used to capture context and higher-order dependencies. The attention matrix $A(\cdot)$ indicates a score between N queries Q and N_k keys, which indicates which part of the input sequence to focus on. $\sigma(\cdot)$ is an activation function (generally $\text{softmax}(\cdot)$).

$$A(Q, K) = \sigma(QK^T) \quad (1)$$

To capture the relations among the input sequence, the values V are weighted by the scores from Equation 1. Therefore, we have

$$\text{SelfAttention}(Q, K, V) = A(Q, K) \cdot V \quad (2)$$

A key property of the self-attention model described above is that it is equivariant to the input order, i.e. it gives the same output independent of how the N input tokens are shuffled. As mentioned above, this is the primary motivation for our work.

2.2 View-based Shape Classification and Retrieval

Point clouds in their raw XYZ format are irregular, unordered and therefore researchers have tried to convert these point clouds into more regular representations like voxels, images. These view images are created by projecting a point cloud onto various planes around it and they are passed onto a image feature extractor to derive view features. These view features are then pooled to create the final shape descriptor used to predict classes. However, existing methods do not consider the relative importance of each view which can affect the performance of the model (Eg: given 10 views for a roof point cloud, there might exist some views which contain

occlusions, missing regions, roof structures projected in a deformed manner onto a single plane which can all confuse the model).

Therefore in this work, we primarily focus on two problems: evaluating if deep learning is suitable for ALS point cloud representation, incorporating attention to derive view-wise feature importance, and create the final shape descriptor dynamically. This is done using a simplified attention operation that derives only K and V vectors from the view features. The obtained K vectors are softmax normalized to derive view-wise importance values which are then multiplied with the V and added to derive the final shape representation. This helps derive richer shape descriptors which are reflected in the improved shape classification and retrieval performance discussed in subsection. 3.1.

2.3 Point-based Shape Classification and Retrieval

In this work, we would like to explore and confirm our hypothesis that attention can act as a core operator in point cloud representation methods by proposing a fully attentional network. As discussed in subsection. 2.1, in NLP each element in the sequence corresponds to a word and the same idea is applicable for a sequence of N discrete objects, like points in a point cloud. However, due to a large number of elements in this sequence (>1000 points in a point cloud), naively applying attention can lead to over-parametrization or sparse attention matrices. Additionally, existing works in NLP like [13] have shown that directly applying pooling on the derived point-wise features after attention can lead to poor performance.

Therefore, we propose a novel iterative transformer that shares parameters across multiple blocks, and only the K , V vectors are updated while the Q vector is kept the same. This leads to a sequential feature learning procedure similar to an unfolded Recurrent Neural Network. It helps capture hierarchical features without over-parametrizing the network. Additionally, we introduce a dynamic grouping module that routes the points into corresponding region-specific groups (in a learnable fashion) and creates group-wise feature vectors. These group-wise feature vectors are then aggregated

to create the final shape descriptor used for classification and retrieval. The proposed method showcases improved performance in both synthetic, real benchmarks with much fewer parameters. The learnable nature of all the operations leads to improved robustness and is evaluated and discussed in subsection. 3.2.

2.4 Point-based Shape Completion

Inspired by the success of our previous work PT, we attempt to solve a more complex problem of shape completion using a fully-attention network. Most existing methods use a standard point-based backbone (eg: PointNet, PointNet++) and derive global features which are then used to create the final point cloud using a series of folding blocks and/or MLPs. However, they fail to retain local geometric information, i.e given four distinctly unique chairs in the dataset, autoencoders tend to "average" these shapes and produce a common structure that can minimize loss against all the samples. Further, they do not reconstruct corners, edges, thin lines effectively and produce a lot of noise. This can be problematic in the case of roofs as they are primarily composed of such simple geometric shapes and the presence of noise/outliers can confuse these models to think that there exists some shape in these regions.

Therefore, we propose a novel architecture first by modifying attention to a strictly local operation by forcing each point to focus only on its immediate neighborhood. This helps derive stronger features describing each local geometry when compared to other methods. The partial point cloud is sequentially downsampled and multi-resolution features are extracted using a stack of local attention blocks in the encoder. The final extracted feature is used to derive a coarse complete point cloud which undergoes a series of up-sampling operations in the decoder. After each up-sampling operation, the generated point cloud is corrected using cues derived from the encoder which is used to guide the decoder to reconstruct a complete point cloud coherent to the input partial shape. Finally, to uniformly redistribute the points obtained in the complete point cloud, a refinement block is added to get the final output. The results are discussed in brief in subsection. 3.3.

3 Results and Discussion

In this section, we highlight key results and inferences from the proposed methods. Due to space constraints, we cannot describe every result in detail in this report and urge the readers to read the corresponding papers associated with the proposed methods. For the shape classification, retrieval experiments, we use the RoofN3D dataset: contains roof instances from a large-scale urban ALS scan of New York city and popular synthetic benchmarks ModelNet40, RobustPointSet. For the shape completion experiments, we use a

damaged set derived from the RoofN3D dataset and synthetic benchmark Completion3D.

3.1 View-based Shape Classification and Retrieval

Due to no prior work in view-based shape classification in ALS roof point clouds, we first set up a baseline using a single view ResNet classifier, followed by a multi-view ResNet classifier. Then we incorporate our proposed method and it achieves 1.24% higher classification scores and 6.14 higher Mean Average Precision (MAP) in shape retrieval when compared to existing methods in the RoofN3D dataset. The higher retrieval scores indicate that the extracted features are more descriptive. To empirically validate the effectiveness of our view-wise pooling technique we compare with other naive strategies like max-pooling and mean-pooling. Further, we try to mimic these naive pooling strategies using our view-wise pooling methods by including entropy and produce diffuse (mean-pooling) and concentrated (max-pooling) importance weights. Both these experiments yield lower performance which clearly indicates that learning view-wise relevance is important to derive a better shape representation.

3.2 Point-based Shape Classification and Retrieval

Since we are exploring a new design space i.e fully attentional model, we also evaluate the performance of the proposed method on the standard benchmark dataset ModelNet40. PT achieves 1.26% higher accuracy in the roof classification task and 6.54 higher MAP scores when compared to existing methods. It is important to note that RoofN3D has a limited number of classes and is highly imbalanced (since it's a real dataset) and therefore it is expected that there won't be a major improvement from MVCNN-SA to PT in the RoofN3D dataset. However, in the benchmark ModelNet40 dataset, PT achieves 92.5% accuracy in the shape classification task, 88.4 MAP in the retrieval task, and performed better than SOTA graph-based and other attention-based methods. Additionally, to evaluate robustness, we perform detailed experiments on the RobustPointSet dataset where the model is tested on unseen input corruptions and PT stood out as the best performing method achieving 62.5% average accuracy. We perform a similar set of experiments in RoofN3D and PT outperformed existing methods by 3.57% in average accuracy across various corruptions in roof classification while maintaining a 90.77 MAP score in partial shape roof retrieval.

3.3 Point-based Shape Completion

In this work, to fully evaluate the model, we test the model on both RoofN3D and benchmark dataset Completion3D. PCT showcased better performance by achieving 24×10^4 chamfer distance in roof completion and 12×10^4 in Completion3D. We want to point out that the given metric - chamfer

distance is not ideal to evaluate the quality of reconstruction. Chamfer distance does not penalize the model based on its ability to retain local geometric information, and it is sufficient to predict the overall geometry. We strongly believe this is one of the reasons our method does not beat SOTA methods in Completion3D while still producing better visual results. However, we could have used more accurate metrics like Earth Movers Distance (EMD) but due to computational challenges and its behavior to approximate at higher resolutions, we choose not to (refer to known bugs in [issue](#)).

4 Future Work

- In this thesis, we explored attention - as a complimentary sub-layer or as a core operation in a fully-attention network in three tasks - classification, retrieval, and shape completion. However, its complete effectiveness is not yet explored for various other tasks - like segmentation, noise removal, etc.
- Additionally, this method can also be extended for larger point sets - aerial city scene, driving lidar scans, etc., for tasks applicable for autonomous vehicles.
- Our proposed methods were seen to be inherently robust to various unseen transformations, and this can be further improved, especially in the case of rotation, by introducing an orientation invariant input representation (like distance and polar angles).
- One of the major bottlenecks in attention is the scalar dot product operation which has an $O(N^2)$ time complexity, and reducing the same to lower orders like $O(N)$ is still an open research problem.

5 Conclusion

This thesis is perhaps the foremost research for applying DL to derive efficient representations for ALS roof point clouds. We also highlight areas that are less looked upon like the choice of pooling operations to derive shape descriptors and how important modifications in these operations lead to improved performance. We hypothesize and showcase that attention can be a better operator to derive point cloud representations due to their inherent set-like properties. We explore new design considerations like sharing parameters across layers and only updating some of the outputs of each layer. This raises important questions - do we actually need multiple independent layers in large transformer models which is extremely relevant not just in the case of point clouds but in NLP and other vision topics as well. We also showcase how simple modifications to attention can convert it from a global to a more locally activated operation. This can help solve more challenging problems which require such local biases like generation, detection, segmentation, etc. The proposed networks show strong performance on synthetic benchmarks and real datasets, which need not be the case always. We hope this inspires future works to

also evaluate using similar strategies to fully understand the effectiveness of their methods.

6 Publications

The following are the publications which are the outcome of this research work:

1. [Roof Classification from 3D LiDAR Point Clouds using Multi-view CNN with Self-Attention](#)
IEEE Geoscience and Remote Sensing Letters; Sibgrapi 2019
2. [Point Transformer for Shape Classification and Retrieval of Urban Roof Point Clouds](#)
IEEE Geoscience and Remote Sensing Letters
3. Point Completion Transformer for Shape Completion of ALS Roof Point Clouds
In submission

7 Acknowledgments

The work was done in Advanced Geometric Lab of IIT Madras, Chennai, India. I thank my supervisor, co-authors and colleagues in AGCL and IITM for their help, in particular for crucial help with the infrastructure and the useful discussions.

References

- [1] Fatemeh Alidoost and Hossein Arefi. 2016. KNOWLEDGE BASED 3-D BUILDING MODEL RECOGNITION USING CONVOLUTIONAL NEURAL NETWORKS FROM LIDAR AND AERIAL IMAGERIES. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLI-B3 (06 2016), 833–840. <https://doi.org/10.5194/isprs-archives-XLI-B3-833-2016>
- [2] M. Axelsson, U. Soderman, A. Berg, and T. Lithen. 2018. Roof Type Classification Using Deep Convolutional Neural Networks on Low Resolution Photogrammetric Point Clouds From Aerial Imagery. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1293–1297. <https://doi.org/10.1109/ICASSP.2018.8461740>
- [3] D. Chen, R. Wang, and J. Peethambaran. 2017. Topologically Aware Building Rooftop Reconstruction From Airborne Laser Scanning Point Clouds. *IEEE Transactions on Geoscience and Remote Sensing* 55, 12 (2017), 7032–7052. <https://doi.org/10.1109/TGRS.2017.2738439>
- [4] Martin A. Fischler and Robert C. Bolles. 1981. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Commun. ACM* 24, 6 (June 1981), 381–395. <https://doi.org/10.1145/358669.358692>
- [5] André Henn, Gerhard Gröger, Viktor Stroh, and Lutz Plümer. 2013. Model driven reconstruction of roofs from sparse LIDAR point clouds. *ISPRS Journal of Photogrammetry and Remote Sensing* 76 (Feb. 2013), 17–29. <https://doi.org/10.1016/j.isprsjprs.2012.11.004>
- [6] Florent Lafarge and Clément Mallet. 2011. Building large urban environments from unstructured point data. In *2011 International Conference on Computer Vision*. 1068–1075. <https://doi.org/10.1109/ICCV.2011.6126353>
- [7] F. Lafarge and C. Mallet. 2012. Creating Large-Scale City Models from 3-D-Point Clouds: A Robust Approach with Hybrid Representation. *International Journal of Computer Vision* 99 (2012), 69–85.
- [8] Haoming Lu and Humphrey Shi. 2021. Deep Learning for 3-D Point Cloud Understanding: A Survey. *arXiv:2009.08920 [cs.CV]*

- [9] Konstantinos Makantasis, Konstantinos Karantzas, Anastasios Doulamis, and Konstantinos Loupos. 2015. Deep Learning-Based Man-Made Object Detection from Hyperspectral Data. *Lecture Notes in Computer Science* 9474, 717–727. https://doi.org/10.1007/978-3-319-27857-5_64
- [10] P. Musialski, P. Wonka, D. G. Aliaga, M. Wimmer, L. Gool, and W. Purgathofer. 2013. A Survey of Urban Reconstruction. *Comput. Graph. Forum* 32, 6 (Sept. 2013), 146–177. <https://doi.org/10.1111/cgf.12077>
- [11] M. Omidalizand and M. Saadatseresht. 2013. SEGMENTATION AND CLASSIFICATION OF POINT CLOUDS FROM DENSE AERIAL IMAGE MATCHING. *The International Journal of Multimedia & Its Applications* 5 (2013), 33–51.
- [12] Carlos A. Vanegas, Daniel G. Aliaga, and Bedrich Benes. 2012. Automatic Extraction of Manhattan-World Building Masses from 3-D Laser Range Scans. *IEEE Transactions on Visualization and Computer Graphics* 18, 10 (2012), 1627–1637. <https://doi.org/10.1109/TVCG.2012.30>
- [13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems* 30, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 5998–6008. <http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf>
- [14] V. Verma, R. Kumar, and S. Hsu. 2006. 3-D Building Detection and Modeling from Aerial LIDAR Data. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, Vol. 2. 2213–2220. <https://doi.org/10.1109/CVPR.2006.12>
- [15] George Vosselman and Hans-Gerd Maas. 2010. *Airborne and Terrestrial Laser Scanning*. Whittles Publishing. I–XVII, 1–318 pages.
- [16] R. Wang, J. Peethambaran, and D. Chen. 2018. LiDAR Point Clouds to 3-D Urban Models: A Review. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 11, 2 (2018), 606–627. <https://doi.org/10.1109/JSTARS.2017.2781132>
- [17] Andreas Wichmann. 2018. *Grammar-guided reconstruction of semantic 3-D building models from airborne LiDAR data using half-space modeling*. Doctoral Thesis. Technische Universität Berlin, Berlin. <https://doi.org/10.14279/depositonce-6803>
- [18] Jixing Yan, Jie Shan, and Wanshou Jiang. 2014. A global optimization approach to roof segmentation from airborne lidar point clouds. *ISPRS Journal of Photogrammetry and Remote Sensing* 94 (2014), 183–193. <https://doi.org/10.1016/j.isprsjprs.2014.04.022>
- [19] Xi Zhang, Andi Zang, Gady Agam, and Xin Chen. 2014. Learning from Synthetic Models for Roof Style Classification in Point Clouds. In *Proceedings of the 22Nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (Dallas, Texas) (SIGSPATIAL '14)*. ACM, New York, NY, USA, 263–270. <https://doi.org/10.1145/2666310.2666407>
- [20] Qian-Yi Zhou and Ulrich Neumann. 2008. Fast and Extensible Building Modeling from Airborne LiDAR Data. In *Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (Irvine, California) (GIS '08)*. Association for Computing Machinery, New York, NY, USA, Article 7, 8 pages. <https://doi.org/10.1145/1463434.1463444>