Limitations and Review of Geometric Deep Learning Algorithms for Monocular 3D Reconstruction in Architecture

Alberto Tono
Meher Shashwat Nigam
Stasya Fedorova
Amirhossein Ahmadnia
Cecilia Bolognesi

Abstract

This paper aims to test algorithms for 3D reconstruction from a single image specifically for building envelopes. This research shows the current limitations of these approaches when applied to classes outside of the initial distribution. We tested solutions with differentiable rendering, implicit functions, and other end–to–end geometric deep learning approaches. We recognize the importance of generating a 3D reconstruction from a single image for many different industries, not only for Architecture, Engineering, and Construction (AEC) industry but also for robotics, autonomous driving, gaming, virtual and augmented reality, drone delivery, 3D authoring, improving 2D recognition and many others. Henceforth, engineers and computer scientists could benefit, not only from having the 3D representations but also from the Building Information Model (BIM) at their disposal. With further development of these algorithms it could be possible to access specific properties such as thermal, physical, maintenance, cost, and other parameters embedded in the class.

Keywords
geometric deep learning, monocular 3D reconstruction, building envelope, architecture.

doi.org/10.3280/oa-686.68

Introduction

Currently, we mainly capture reality around us with static 2D media, such as pictures, or videos, in case we add the temporal component to it. For instance, architects represent their work with iconic pictures or render that convey their styles [Yoshimura et al. 2019], losing the sense of immersion provided by volumetric representations that can allow the user to explore the environment thanks to real–time rendering. Furthermore, when architects perform surveys, they capture the environment with methods that are prone to errors, motionless pictures, expensive laser scanners, or other methodologies for classifying and streamline workflows [Grilli et al. 2019; Matrone et al. 2020; Xia et al. 2018; Chang et al. 2017]. Today technologies aim to create a more immersive experience that can help in the long term to fill the gap between a 2D representation and the 3D physical space (in this paper, we won't consider temporal representations and others related to higher dimensional spaces [Rempe et al. 2020]). Thanks to depth sensors, lidar sensors, stereo imagery, it is possible to capture more information that helps us obtain 3D representations from 2D media like videos, panorama pictures, or even single 2D pictures. State–of–the–art (SOTA) algorithms are democratizing how we generate 3D objects from a multi–view or single view representation. For example, the so–called 3D photos produce a more immersive and dynamic representation [Kopf et al. 2020], allowing users and consumers to interact with their media thanks to the engineering use of the gyroscope in the devices.
The multidisciplinary inherited advancements in these technologies will provide better machine perception, a more immersive environment, and instant geometrical representations of objects and space [Keshavarzi et al. 2019; McCormac et al. 2016]. For example, nowadays, AR/VR experiences require an initial calibration process for the headsets. This is not instantaneous and requires an accurate scanning of the environment creating an adoption barrier for new users. Allowing an instantaneous representation of the environment from a single picture can benefit many applications, not only for AR/VR, but also indoor robot/drone navigation, especially within the building environment, where the environment is dynamic and subject to continue transformations. Such methods will allow easy authoring of 3D content, users will be able to obtain the 3D reconstructions of objects after taking a picture. The obtained reconstruction could be modiufied further as desired and would serve as a good, realistic starting point saving lot of effort. After presenting the importance of converting a monocular image instantly into a 3D model, we need to analyze the output formats produced: the file format [Ahmed et al. 2018], geometric representation, and dataset format [Gao et al. 2020] . Approaches like Mesh–RCNN [Gkioxari et al. 2019] produce 3D meshes by first identifying the objects in the image (Faster RCNN/ MaskRCNN [Gkioxari et al. 2019; Ren et al. 2017; Girshick 2015]) and then predicting coarse voxelized object, which is further refined to pr duce meshes. These meshes can later be sampled to point clouds where metrics such as chamfer distance and EMD can be applied. Other procedural methods have been taken into consideration and examined [Nishida et al. 2018; Liu et al. 2017]. Unfortunately, they lack flexibility, and they require considerable efforts during the initial stages to define a shape grammar that can produce the desired output. In this research, we tested and compared different approaches explaining their potential and current limitations in the Architectural Heritage. We tested: Mesh–RCNN, (figs. 0-1) Occupancy Networks [Mescheder et al. 2019], Pix2Mesh [Wang et al. 2018] and other solutions into the wild. These AI–powered techniques can blend digital and reality in a much more democratic way without expensive and bulky HMDs with multiple cameras. This paper experiments with new functional differentiable rendering frameworks like Pytorch3D (used in MeshRCNN) to explore 2D–3D neural networks. Moreover, working with 3d embedded semantics [Zhang et al., 2020], hierarchical graph network [Chen et al. 2020], it could be possible to encode shapes into images and learning their 3D part assembly from a single image [Li et al. 2020]. For example, after taking a picture of a façade, it would be possible to recognize its parts and regenerate a 3D model with windows, doors, balconies, and other sub–parts with associated information (BIM), and semantic properties ontologies. In this paper, an extensive review of state–of–the–art methods is presented to better understand current limitations and opportunities specifically for architecture.

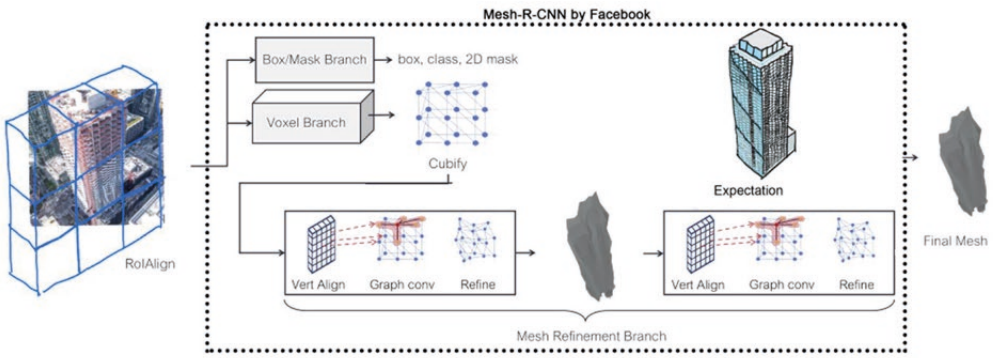

Fig. 1. Original from Facebook MeshRCNN – Adaptation to Architectural Field. (Testing Mesh–RCNN on the pictures of building envelopes).

Related Work

Methods for 3D reconstruction from single–image are complicated by the fact there could be many possible reconstructions when the object is not entirely visible; hence, most of them need to rely on strong supervision. Therefore, they use datasets such as ShapeNet or ModelNet [Wu et al. 2015] Other methods learn from images paired with aligned 3D meshes or require keypoint annotations on the 2D training images [Wu et al. 2016] and/or multiple views for each object instance, often with pose annotations. Shading becomes an important cue for 3D understanding, explored in numerous works over the years [Henderson et al. 2020]. Different methods have been explored in the past: mesh based such as N3MR [Kato et al. 2018], or voxel based like 3D–R2N2 [Choy et al. 2016] and MVD [Smith et al. 2018], or point based like PSG [Fan et al. 2017] and many others [Aubry et al. 2014]. These have issues in performing a complete task with objects not within the training distribution, so we wanted to confirm our hypothesis and stress these limitations [Henderson et al. 2020; Wang et al. 2019].

3D Reconstruction From a Single Image

Learning–based 3D reconstruction works are based on different 3D representations as presented before. While voxel representations prove to be computationally expensive, point cloud representations are demonstrated to be rotation and translation invariant, and computationally more efficient than voxels [Liu et al. 2019]. Moreover, mesh representations, better preserve the connections between distinct parts and are more suitable for fine–grain detailed representations. Modern implicit functions not only prove to be extremely efficient with their continuous and differentiable representation of the iso–surface with a binary value indicating whether a point is within the volume, but also more accurate for tasks such as reconstruction and 3D shape completion [Gu et al. 2020]. Nerf, Occupancy Network, DeftTef [Gao et al. 2020] have recently followed for this task.
Within the AEC, 3D shapes and objects preserve a common grammar and they are composed by a fixed set of components such as windows, doors, roof, floors, walls, and others. While the typology can change, the main elements in the building stay the same for most cases (except for some iconic buildings and pavilions). The philosophy of Hoffman and Richards influenced this research. In fact, they viewed object recognition tasks as a visual system decomposition of shapes into parts with their descriptions and spatial relations. In the same way, we propose that the best way of representing a building reconstruction is to assemble each component together, orienting their quaternions to perfectly fit an initial picture which was inspired by the CompoNet work [Schor et al. 2019]. In contrast to the approach, we aim to translate the assembly algorithm, specifically for an an architectural task. They used a generative neural network for generating 3D shapes from a 2D image, based on a part–based prior, where the key idea was for the network to synthesize shapes by varying both the shape parts and their compositions. Treating a shape not as an unstructured whole, but as a composable set of deformable parts, adds a combinatorial dimension to the generative process to enrich the diversity of the output, encouraging the generator to venture more into the "unseen".

They generated a plethora of shapes compared with baseline generative models using their custom metrics. The assembly–based synthesis was inspired by 3D shape assembly research that generates new shapes from a combination of various parts [Huang et al. 2020; Li et al. 2019] from a single image.

Conclusion

We saw that projects such as Mesh–RCNN lack the ability to perform well with unseen classes. This limitation of generalizing to unseen classes make these approaches challenging to adopt. Furthermore, the training of these algorithms required multi–GPU training(8 GPUs V100, for Mesh–RCNN) that not all the researchers can access. The current lack of a common balanced dataset (with intra and inter–class variance), or pre–trained models that generalize well to unseen data, are missing in the research community, and with this research we hope to stress the importance of the creation of such datasets and models. Another limitation is embedded in the metrics used to evaluate the performance of these algorithms: chamfer distance, EMD (earth moving distance), mAP and others offer good quantitative results distant from a recognizable representation that follows qualitative results. Finally, the creation of such dataset could provide new research on 3D shape explorations for architects using Generative Adversarial Network in 3D like ShapeGAN and 3DGAN [Kleineberg et al. 2020; Freeman et al. 2016].

Acknowledgment

We would like to thank Andrea Giordano, UniPd, Reaach–Id conference for the opportunity to publish our work. Georgia Gkioxari, Kaichun Mo, Silvio Savarese, Martin Fischer, Andrea Tagliasacchi, Nicolas Chaulet, Lamberto Ballan, Dmitry Kudinov, Mohammed Keshavari, Andrean Zani, Ignacio Garcia Dorado for valuable discussions. We would also like to thank the Computational Design Institute.

References

Ahmed Eman, Saint Alexandre, Shabayek Abd El Rahman, Cherenkova Kseniya, Das Rig, Gusev Gleb, Aouada Djamila, Ottersten Bjorn (2018). A survey on deep learning advances on different 3D data representations. In *arXiv*, 1 (1), pp. 1-35.

Aubry Mathieu, Maturana Daniel, Efros Alexei A., Russell Bryan C., Sivic Josef (2014). Seeing 3D chairs: Exemplar part–based 2D–3D alignment using a large dataset of CAD models. In *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 3762-3769.

Chang Angel, Dai Angela, Funkhouser Thomas, Halber Maciej, Nießner Matthias, Savva Manolis, Song Shuran, Zeng Andy, Zhang Yinda (2018). Matterport3D: Learning from RGB–D data in indoor environments. In *Proc. – 2017 Int. Conf. 3D Vision, 3DV 2017*, pp. 667-676.

Chen Jintai, Lei Biwen, Song Qingyu, Ying Haochao, Chen Danny Z. , Wu Jian (2020). A Hierarchical Graph Network for 3D Object Detection on Point Clouds. In *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 389-398.

Choy Christopher B., Xu Danfei, Gwak JunYoung, Chen Kevin, Savarese Silvio (2016). 3D–R2N2: A unified approach for single and multi–view 3D object reconstruction. In *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics),* vol. 9912 LNCS, pp. 628-644.

Fan Haoqiang, Su Hao, Guibas Leonidas J. (2017). A point set generation network for 3D object reconstruction from a single image. In *Proc. – 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017,* pp. 2463-2471.

Gao Jun, Chen Wenzheng, Xiang Tommy, Jacobson Alec, McGuire Morgan, Fidler Sanja (2020). Learning deformable tetrahedral meshes for 3D reconstruction. In *arXiv*, *NeurIPS*, pp. 1-12.

Girshick Ross (2015). Fast R–CNN. In *Proc. IEEE Int. Conf. Comput. Vis., ICCV,* pp. 1440-1448.

Gkioxari Georgia, Malik Jitendra, Johnson Justin (2019) Mesh R–CNN. In *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 9784-9794.

Grilli Eleonora, Remondino Fabio (2019). Classification of 3D digital heritage. In *Remote Sensing,* 11(7), pp. 1-23.

Gu Jiayuan, Ma Wei–Chiu, Manivasagam Sivabalan, Zeng Wenyuan, Wang Zihao, Xiong Yuwen, Su Hao, Urtasun Raquel (2020). Weakly–Supervised 3D Shape Completion in the Wild. In *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics),* 12350 LNCS, pp. 283-299.

Henderson Paul, Ferrari Vittorio (2020). Learning Single–Image 3D Reconstruction by Generative Modelling of Shape, Pose and Shading. In *Int. J. Comput. Vis.*, 128 (4), pp. 835-854.

Huang Jialei, Zhan Guanqi, Fan Qingnan, Mo Kaichun, Shao Lin, Chen Baoquan, Guibas Leonidas, Dong Hao (2020). Generative 3D Part Assembly via Dynamic Graph Learning. In *arXiv*, NeurIPS, pp. 1-19.

Kato Hiroharu, Ushiku Yoshitaka, Harada Tatsuya (2018). Neural 3D Mesh Renderer. In *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 3907-3916.

Keshavarzi Mohammad, Wu Michael, Chin Michael N., Chin Robert N., Yang Allen Y. (2018). Affordance analysis of virtual and augmented reality mediated communication. In *arXiv*, pp. 1-15.

Kleineberg Marian, Fey Matthias, Weichert Frank (2020). Adversarial generation of continuous implicit shape representations. In *arXiv,* pp. 1-6.

Kopf Johannes et al. (2020). One Shot 3D Photography, In *ACM Trans. Graph*., 39 (4), 76, pp. 1-13 .

Li Jun, Niu Chengjie, Xu Kai (2019). Learning Part Generation and Assembly for Structure–aware Shape Synthesis. In *AAAI Technical Track: Vision,* 34 (07), pp. 1-8.

Li Yichen, Mo Kaichun, Shao Lin, Sung Minhyuk, Guibas Leonidas (2020). Learning 3D Part Assembly from a Single Image. In *Lecture Notes in Computer Science*, 12351, pp. 1-25.

Liu Hantang, Zhang Jialiang, Zhu Jianke, Hoi Steven C. H. (2017). Deepfacade: A deep learning approach to facade parsing. In *IJCAI Int. Jt. Conf. Artif. Intell.,* 0, pp. 2301-2307.

Liu Zhijian, Tang Haotian, Lin Yujun, Han Song (2019). Point–voxel CNN for efficient 3D deep learning. In *Adv. Neural Inf. Process. Syst.,* vol. 32, no. NeurIPS, pp. 1-11.

Matrone Francesca, Grilli Eleonora, Martini Massimo, Paolanti Marina, Pierdicca Roberto, Remondino Fabio (2020). Comparing machine and deep learning methods for large 3D heritage semantic segmentation. In *ISPRS Int. J. Geo–Information*, 9 (9), pp. 1-22.

McCormac John, Handa Ankur, Leutenegger Stefan, Davison Andrew J. (2016). SceneNet RGB–D: 5M Photorealistic Images of Synthetic Indoor Trajectories with Ground Truth. In *arXiv,* pp. 1-11.

Mescheder Lars, Oechsle Michael, Niemeyer Michael, Nowozin Sebastian, Geiger Andreas (2019). Occupancy networks: Learning 3D reconstruction in function space. In *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.,* pp. 4455-4465.

Nishida Gen, Bousseau Adrien, Aliaga Daniel G. (2018). Procedural modeling of a building from a single image. In *Comput. Graph. Forum*, 37 (2), pp. 415-429.

Rempe Davis, Birdal Tolga, Zhao Yongheng, Gojcic Zan, Sridhar Srinath, Guibas Leonidas J. (2020). CaSPR: Learning canonical spatiotemporal point cloud representations. In *arXiv*, NeurIPS, pp. 1-28.

Ren Shaoqing, He Kaiming, Girshick Ross, Sun Jian (2017). Faster R–CNN: Towards Real–Time Object Detection with Region Proposal Networks. In *IEEE Trans. Pattern Anal. Mach. Intell*., 39 (6), pp. 1137-1149.

Schor Nadav, Katzir Oren, Zhang Hao, Cohen–Or Daniel (2019). CompoNet: Learning to generate the unseen by part synthesis and composition. In *Proc. IEEE Int. Conf. Comput. Vis.,* pp. 8758-8767.

Smith Edward, Fujimoto Scott, Meger David (2018). Multi–view silhouette and depth decomposition for high resolution 3D object representation. In *Adv. Neural Inf. Process. Syst.,* NeurIPS, pp. 6478-6488.

Xu Qiangeng, Wang Weiyue, Ceylan Duygu, Mech Radomir, Neumann Ulrich (2019). DISN: Deep implicit surface network for high–quality singleview 3d reconstruction. In *arXiv*, CD, pp. 1-11.

Wang Nanyang, Zhang Yinda, Li Zhuwen, Fu Yanwei, Liu Wei, Jiang Yu–Gang (2018). Pixel2Mesh – Generating Meshes from Single RGB Images. In *Eccv,* p. 1-16.

Wu Jiajun et al. (2016). Single image 3D interpreter network. In *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, 9910 LNCS, pp. 365-382.

Wu Jiajun, Zhang Chengkai, Xue Tianfan, Tenenbaum Joshua B., Freeman William T. (2016). Learning a Probabilistic Latent Space of Object Shapes via 3D Generative–Adversarial Modeling. In *30th Conference on Neural Information Processing Systems*, pp. 1-9.

Wu Zhirong et al. (2015). 3D ShapeNets: A deep representation for volumetric shapes. In *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.,* pp. 1912-1920.

Xia Fei, Zamir Amir R., He Zhiyang, Sax Alexander, Malik Jitendra, Savarese Silvio (2018). Gibson Env: Real–World Perception for Embodied Agents. In *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.,* pp. 9068-9079.

Yoshimura Yuji, Cai Bill, Wang Zhoutong, Ratti Carlo (2019). Deep learning architect: Classification for architectural design through the eye of artificial intelligence. In *Lect. Notes Geoinf. Cartogr.,* pp. 249-265.

Zhang Dongsu, Chun Junha, Cha Sang Kyun, Kim Young Min (2020). Spatial Semantic Embedding Network: Fast 3D Instance Segmentation with Deep Metric Learning. In *arXiv,* pp. 1-8.

Authors
Alberto Tono, CDI, Computational Design Institude, San Francisco, United States, alberto.tono@cd.institute
Meher Shashwat Nigam, IIIT, International Institute of Information Technology, Hyderabad, India, meher.shashwat@students.iiit.ac.in
Stasya Fedorova, Dept. of Architecture, Built environment and Construction engineering, Politecnico di Milano, stanislava.fedorova@mail.polimi.it
Amirhossein Ahmadnia, Dept. of Architecture, Built environment and Construction engineering, Politecnico di Milano, amirhossein.Ahmadnia@polimi.it
Cecilia Bolognesi, Dept. of Architecture, Built environment and Construction engineering, Politecnico di Milano, cecilia.bolognesi@polimi.it

 Isbn 9788835125280

Augmented Reality (AR) and Artificial Intelligence (AI) are technological domains that closely interact with space at architectural and urban scale in the broader ambits of cultural heritage and innovative design. The growing interest is perceivable in many fields of knowledge, supported by the rapid development and advancement of theory and application, software and devices, fueling a pervasive phenomenon within our daily lives. These technologies demonstrate to be best exploited when their application and other information and communication technology (ICT) advancements achieve a continuum. In particular, AR defines an alternative path to observe, analyze and communicate space and artifacts. Besides, AI opens future scenarios in data processing, redefining the relationship between man and computer.
In the last few years, the AR/AI expansion and relationship have raised deep trans–disciplinary speculation. The research experiences have shown many cross–relations in Architecture and Design domains. Representation studies could arise an international debate as a convergence place of multidisciplinary theoretical and applicative contributions related to architecture, city, environment, tangible and intangible Cultural Heritage.
This book collects 66 papers and identify eight lines of research that may guide future developments.

Andrea Giordano *Dept. of Civil, Environmental and Architectural Engineering, University of Padua*
Michele Russo *Dept. of History, Representation and Restoration of Architecture, Sapienza University of Rome*
Roberta Spallone *Dept. of Architecture and Design, Politecnico di Torino*

FrancoAngeli
La passione per le conoscenze