Building 3D scenes from 2D images: Literature review and Exploration

Jiaming Yu
Boston University
jiamingy@bu.edu

***Abstract*— Building 3D scenes from 2D images is a challenging and significant problem in computer vision and other imaging applications. The 3D reconstruction methods has many types like single-view vs multi-view or conventional methods vs methods using deep net architecture. For the conventional methods, One main type is to analyze the shade information of the image while another type is to utilize the 3D geometric information of the images. There are four types of 3D representation: depth image, voxel, point cloud and mesh and then this paper review the related literature. Also, we would discuss about drone based 3D reconstruction focusing on path planning algorithms: mainly Next-Best-View (NBV) and (Explore-then-Exploit (ETE). In this paper, we also explore the user story of this technique. It has many applications like motion capture and face tracking, 3D scanner, drone 3D mapping and so on. Also, in this paper, we analyze the current situation and future work.**

I. INTRODUCTION

Upon seeing an image, a human would have no difficulty understanding its 3D structure. However, for computer vision systems, recognizing 3D structures would be quite challenging. Understanding 3D structure is a fundamental problem of computer vision and since the establishment of computer vision as a field several decades ago, 3D geometric shape has been considered to be one of the most important cues in this field. Therefore, building 3D scenes from 2D images is a very important problem in computer vision and other imaging applications.

Building 3D scenes from 2D images can be divided into many types according to different aspects. According to the number of views of images, it can be divided into single-view and multi-view methods, or we can also call them Monocular cues methods and Binocular stereo vision. Monocular cues methods refer to using one or more images from one viewpoint (camera) to proceed to 3D construction while Binocular Stereo Vision obtains the 3-dimensional geometric information of an object from multiple images. Obviously, single-view method is more difficult to research.

Besides, there is another way to divide 3D reconstruction approaches into conventional methods and methods relying on deep network architecture.

Conventional 3D reconstruction methods develops under the era of pre deep learning and thus people would rely on mathematical methods. One type of approaches typically use images to obtain depths in the scene and reconstruct a 3D scene using geometric entities, such as points and polygons. Another type of methods is to analyze the shade information in the image like shape-from-shade algorithm. There are many other mathematical approaches and mostly they can be called shape-from-X.

Recently, due to the development of deep learning in computer vision field, there are many methods using a machine learning model. With the advances in deep representation learning and the introduction of large 3D CAD datasets like ShapeNet, there have been some inspiring attempts in learning deep object representations based on voxelized objects. This is a challenging problem because, compared to the space of 2D images, it is more difficult to model the space of 3D shapes due to its higher dimensionality. [12]

A. Related Work

1) Conventional methods: There are mainly two types of conventional 3D reconstruction methods. One is to analyze the shade information of the image. The earliest method of this type is shape-from-shading. In [5], Professor BKP Horn presented a method for finding the shape of a smooth apaque object form a monocular image, given a knowledge of the surface photometry, the position of the lightsource and certain auxiliary information to resolve ambiguities.

Another type is to utilize the 3D geometric information of the images, like Bundle Adjustment algorithm. In [2], Andrew Davison presented a general method as shown in Figure 1 for real-time, vision-only single-camera simultaneous localisation and mapping (SLAM) - an algorithm which is applicable to the localisation of any camera moving through a scene - and study its application to the localisation of a wearable robot with active vision. He utilize a single-view camera als well as geometric information to present the method of Visual SLAM.

Some other methods are also researched. In [9], a method of Markov Random Field (MRF) to infer a set of "plane parameters" that capture both the 3-d location and 3-d orientation of the patch is presented.

Conventional methods are often called Shape-from-X. Nowadays, they are still very popular. For example, in CVPR 2019 best paper, Shumian Xin's team present a novel theory of Fermat paths of light between a known visible scene and an unknown object not in the line of sight of a transient camera. Based on this theory, they present an algorithm, called Fermat Flow, to estimate the shape of the non-line-of-sight object. Our method allows, for the first time, accurate shape recovery of complex objects, ranging from diffuse to specular, that are hidden around the corner as well as hidden behind a diffuser. Finally, their approach is agnostic to the particular technology used for transient imaging. [14]

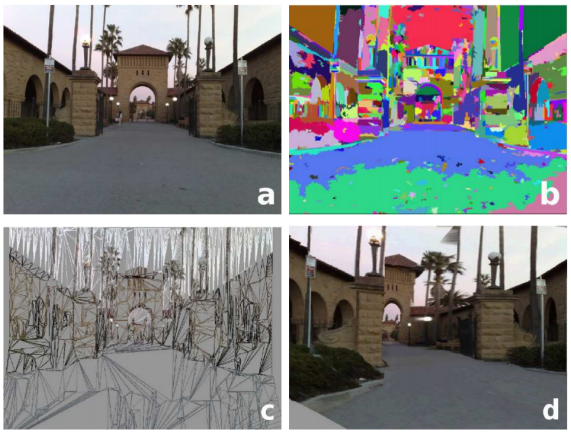

Fig. 1: (a) An original image. (b) Oversegmentation of the image to obtain "superpixels". (c) The 3-d model predicted by the algorithm. (d) A screenshot of the textured 3-d model. [2]

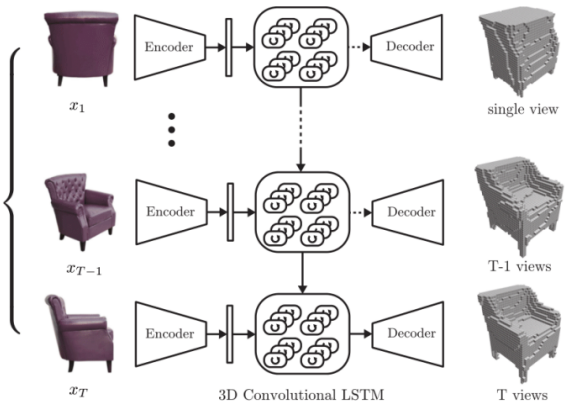

Fig. 2: 3D-R2N2 network [1]

This paper makes a new breakthrough in the traditional 3D reconstruction algorithm.

2) ShapeNet: In [13], Wu presented a model, 3D ShapeNets,learns the distribution of complex 3D shapes across different object categories and arbitrary poses from raw CAD data, and discovers hierarchical compositional part representations automatically. It naturally supports joint object recognition and shape completion from 2.5D depth maps, and it enables active object recognition through view planning. To train their 3D deep learning model, they construct ModelNet - a large-scale 3D CAD model dataset. Extensive experiments show that our 3D deep representation enables significant performance improvement over the-state-of-the-arts in a variety of tasks. It is a popular work to push people to research on the 3D reconstruction methods using deep net architecture.

B. Paper Organization

The contribution of this paper is to literature review 3D reconstruction technique: conventional methods are discussed related work and the methods using deep networks are discussed according to the 3D shape representations. Also, this paper explore the user stories and analyze the current situation and future work.

The paper is organized as follows. The introduction of conventional 3D reconstruction is in section 1 related work. The literature review on methods using deep networks would be in section 2 and in section 2, we would also discuss drone based 3D reconstruction techniques. In section 3, we would explore the user stories of 3D reconstruction techniques and discuss current situation and future work on 3D reconstruction.

II. Literature Review

A. 3D shape representation and related literature

In this section, we would introduce the 3D shape representation and related methods using deep learning architecture. There are four types of normal 3D shape representation: depth, point cloud, voxel and mesh and literature review the corresponding articles.

1) depth: Estimating depth is an important component of understanding geometric relations within a scene. A depth map is an image or image channel that contains information relating to the distance of the surfaces of scene objects from a viewpoint. It is related to Z-depth which relates to a convention that the central axis of view of a camera is in the direction of the camera's Z axis. There is much work on the depth estimation based on multi-view images. Recently, people start to research on the method using deep network to predict the depth of the image. For example, in [3], David Eigen's team presented a method that addresses this task by employing two deep network stacks: one that makes a coarse global prediction based on the entire image, and another that refines this prediction locally. They also apply a scale-invariant error to help measure depth relations rather than scale. By leveraging the raw datasets as large sources of training data, Their method achieves state-of-the-art results on both NYU Depth and KITTI, and matches detailed depth boundaries without the need for superpixelation.

2) voxel: A voxel is a unit of graphic information that defines a point in three-dimensional space. Since a pixel (picture element) defines a point in two dimensional space with its x and y coordinates , a third z coordinate is needed. In 3-D space, each of the coordinates is defined in terms of its position, color, and density. In [1], Christopher B.Choy's team use Encoder-3DLSTM-Decoder deep network to create the reconstruction from 2D images to 3D voxel model.They call their model 3D Recurrent Reconstruction Neural Network (3D-R2N2). They use CNNs to encode images into features. A standard feed-forward CNN and a deep residual variation of it. Then they use a new architecture called 3D-Convolutional LSTM as the recurrence module to retain what it has seen and to update the memory when it sees a new image. Each unit of 3D-LSTM receives the same feature vector from the encoder as well as the hidden states from its neighbors as input, and then passes them to the decoder, 3D Deconvolutional Neural Network. This approach can be used for both the single-view

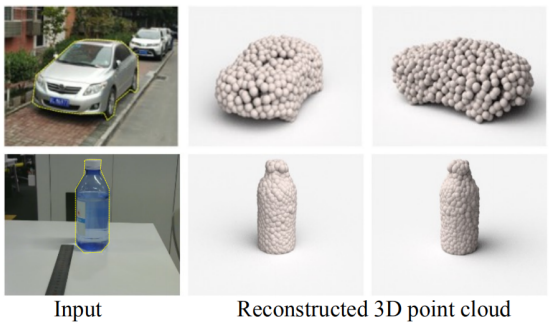

Fig. 3: Point cloud [4]

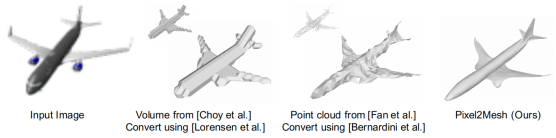

Fig. 4: The method using mesh can achieve higher quality and contains more details [11]

images and multi-view images.

3) point cloud: A point cloud is a simple, uniform structure that is easier to learn, as it does not have to encode multiple primitives or combinatorial connectivity patterns, as is shown in Figure 2. In addition, a point cloud allows simple manipulation when it comes to geometric transformations and deformations, as connectivity does not have to be updated. In [4], Haoqiang Fan's team address the problem of 3D reconstruction from a single image, generating a straightforward form of output point cloud coordinates, that is to say, use the deep net architecture to directly generate a single image to point cloud. Their final solution is a conditional shape sampler, capable of predicting multiple plausible 3D point clouds from an input image.

4) mesh: The estimated 3D shape, as the output of the neural network, is represented as either a volume or point cloud. However, both representations lose important surface details, and is non-trivial to reconstruct a surface model as is shown in Figure 3 while a mesh, which is more desirable for many real applications since it is lightweight, capable of modelling shape details, easy to deform for animation, to name a few. In [11], Nanyang Wang's team proposed an end-to-end deep learning architecture that produces a 3D shape in triangular mesh from a single color image.

B. Drone based 3D reconstruction

Nowadays, with the development of Unmanned Aerial Vehicle (UAV), Drone based 3D reconstruction techniques are also quite emerging research field and there are some special problems to be solved. UAV is a generic aircraft design to operate with no human pilot onboard. UAV platforms are nowadays a valuable source of data for inspection, surveillance, mapping and 3D modeling issues and UAVs can be considered as a low cost alternative to the classical manned aerial photogrammetry. [7]

Compared to normal 3D reconstruction, Drone based 3D reconstruction needs to consider other special problems. Due to the properities of Drone, drone based 3D reconstruction technique would focus on reconstructing large structures and path planning algorithms to ensure a better 3D reconstruction result.

So far there are two main types of path planning algorithms for UAV based 3D reconstruction: Next-Best-View (NBV) and Explore-then-Exploit (ETE).

NBV is essentially a greedy algorithm. Its core idea is to select a point with the highest estimated profit from the reachable target area as the target point according to the conditions of the explored area during each step of the exploration process, and then plan a collision-free Then go to the path with the least cost, update the explored area after arrival, and perform the next round of path planning, and iterate until convergence.

The Explore-then-Exploit algorithm was mainly described in [8]. The whole process can be divided into two major steps: The first step is to obtain a rough and rough 3D model of the target object as the input of the second step. The model either already exists or can be generated on a safe plane. The regular flight trajectory is photographed and reconstructed; the second step is to perform a new round of path planning based on the 3D model output in the first step, and collect information according to the plan, use the information collected in this step to reconstruct and output as the final result.

III. Exploration

A. User stories

Nowadays, 3D reconstruction technology has been widely used in games, movies, mapping, positioning, navigation, autonomous driving, VR/AR, industrial manufacturing, and consumer products. And specially for UAV 3D mapping, it is really useful in archaeology, architecture and many other fields.

1) Motion Capture and Face Tracking: The 3D reconstruction can be used in Motion Capture and Face Tracking, which would be really helpful in game and movie fields. For example, One software called Performer provided by Dynamixyz can obtain the widest and smartest range of markerless facial motion capture solutions to the entertainment market. Using this software, face tracking can be processed from a single view or multi views of an actor while multi view is recommended. As the figure 5 shows, we can capture the facial expression using this technique to build 3D objects from 2D videos. Considering you are a director of a 3D film like Zootopia or a 3D game, this technique would definitely help you a lot. Also, this technique can also be used for movies that require special effects like Marvel Films.

2) Building 3D City from 2D images: 3D City Model after 3D reconstruction as is shown in Figure 6 can be used in surveying and mapping filed. It can also provide the map for navigation and autonomous driving. 3D reconstruction technique, like Visual Simultaneous Localization and Mapping (Visual SLAM) is always used in various robots, such as probes, robots, unmanned aerial vehicles, etc. Autonomous

Fig. 5: The markerless facial motion capture from https://www.dynamixyz.com/

Fig. 6: 3D City Model after 3D reconstruction [6]

vehicles also use Visual SLAM technology to map and understand the surrounding environment. Considering that you are the passenger of an autonomous vehicle, it would be hard for GPS to find the accurate location in both indoor space and metropolitan areas while 3D reconstruction technique like Visual SLAM can help you a lot.

3) 3D scanner: 3D scanners are always designed based on 3D reconstruction approach: Binocular stereo vision method. Under manufacturing and industrial design fields, 3D scanners can be used in 3D printing, CFD (Computational Fluid Dynamics), FEA (Finite Element Analysis) and so on. If you want to design a product, the technique to 3D reconstruct the objects would be really helpful.

4) Buiding 3D scenes for real estate: 3D reconstruction approach can also help you view the apartments or estates. For example, Matterport company as Figure 7 shows create the 3D digital model for people who want to have a 3D virtual tour on their future home.

5) Depth camera for Face ID: 3D reconstruction can also be used to enable Face ID by building 3D models of users' face and head. For example, Apple has created a sophisticated TrueDepth front-facing camera system that uses structured light. Its depth estimation works by having an IR emitter send out 30,000 dots arranged in a regular pattern. And thus

Fig. 7: Matterport from https://matterport.com/

people can rely on Face ID on iPhone X securely.

6) UAV 3D mapping: Specially, Drone based 3D reconstruction technique would be really useful in surveying and mapping field via UAV 3D mapping. Drone captures images of an archeological site, construction site or a 3D real object. With the help of 3D restruction technique, UAV can be employed in Agriculture: producers can take reliable decisions to save money and time; Forestry: assessments of woodlots, fires surveillance, vegetation monitoring, species identification, volume computation as well as silviculture can be accurately performed; Archaeology and architecture: 3D surveying and mapping of sites and man-made structures can be performed with low-altitude image-based approaches; Environment: quick and cheap regular flights allow the monitoring of land and water at multiple epochs; Emergency management: UAV are able to quickly acquire images for the early impact assessment and the rescue planning; Traffic monitoring: surveillance, travel time estimation, trajectories, lane occupancies and incidence response are the most required information and so on [7]

B. Current 3D reconstruction and Future work

Although 3D reconstruction using deep net architecture is very popular recently, the conventional method is still pretty significant. For example, as we discussed before, the CVPR 2019 best paper [14] makes a new breakthrough in the conventional 3D reconstruction methods.

Then for methods using deep networks, convolutional networks for single-view object reconstruction have shown impressive performance and have become a popular subject of research.

However, all existing techniques are united by the idea of having an encoder-decoder network that performs non-trivial reasoning about the 3D structure of the output space. In [10], Maxim Tatarchenko's team set up two alternative approaches that perform image classification and retrieval respectively.

They found that These simple baselines yield better results than state-of-the-art methods,both qualitatively and quantitatively. Therefore, they show that encoder-decoder methods are statistically indistinguishable from these baselines, thus indicating that the current state of the art in single-view object reconstruction does not actually perform reconstruction but image classification.

Therefore, in the future, the methods using deep learning network require much more research and we should not ignore the conventional methods. Both of the methods have their own advantages. Maybe the solution to this problem would be the combination and thus we need to develop both of them.

Besides that, more research on the special application fields like drone based 3D reconstruction is also required in the future.

IV. CONCLUSIONS

Building 3D scenes from 2D images is a challenging and significant problem in computer vision and other imaging applications. The 3D reconstruction methods has many types like single-view vs multi-view or conventional methods vs methods using deep net architecture. For the conventional methods, One main type is to analyze the shade information of the image while another type is to utilize the 3D geometric information of the images. There are four types of 3D representation: depth image, voxel, point cloud and mesh. And many methods using deep networks like convolutional networks based on them are researched. Compared to normal 3D reconstruction, Drone based 3D reconstruction needs to consider other special problems like path planning, so far there are two main solution: NBV algorithm and ETE algorithm. For the user story, 3D reconstruction technique has many applications like motion capture and face tracking, 3D scanner, drone 3D mapping and so on. the methods using deep learning network require much more research due to their weakness and we should not ignore the conventional methods in the future.

REFERENCES

[1] C. B. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *European conference on computer vision*, pages 628–644. Springer, 2016.

[2] A. J. Davison, W. W. Mayol, and D. W. Murray. Real-time localization and mapping with wearable active vision. In *The Second IEEE and ACM International Symposium on Mixed and Augmented Reality, 2003. Proceedings.*, pages 18–27. IEEE, 2003.

[3] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems*, pages 2366–2374, 2014.

[4] H. Fan, H. Su, and L. J. Guibas. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 605–613, 2017.

[5] B. K. Horn. Shape from shading: A method for obtaining the shape of a smooth opaque object from one view. 1970.

[6] A. Koutsoudis, F. Arnaoutoglou, and C. Chamzas. On 3d reconstruction of the old city of xanthi. a minimum budget approach to virtual touring based on photogrammetry. *Journal of Cultural Heritage*, 8(1):26–31, 2007.

[7] F. Nex and F. Remondino. Uav for 3d mapping applications: a review. *Applied geomatics*, 6(1):1–15, 2014.

[8] M. Roberts, D. Dey, A. Truong, S. Sinha, S. Shah, A. Kapoor, P. Hanrahan, and N. Joshi. Submodular trajectory optimization for aerial 3d scanning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5324–5333, 2017.

[9] A. Saxena, M. Sun, and A. Y. Ng. Make3d: Learning 3d scene structure from a single still image. *IEEE transactions on pattern analysis and machine intelligence*, 31(5):824–840, 2008.

[10] M. Tatarchenko, S. R. Richter, R. Ranftl, Z. Li, V. Koltun, and T. Brox. What do single-view 3d reconstruction networks learn? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3405–3414, 2019.

[11] N. Wang, Y. Zhang, Z. Li, Y. Fu, W. Liu, and Y.-G. Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 52–67, 2018.

[12] J. Wu, C. Zhang, T. Xue, B. Freeman, and J. Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *Advances in neural information processing systems*, pages 82–90, 2016.

[13] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015.

[14] S. Xin, S. Nousias, K. N. Kutulakos, A. C. Sankaranarayanan, S. G. Narasimhan, and I. Gkioulekas. A theory of fermat paths for non-line-of-sight shape reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6800–6809, 2019.