2021 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct) | 978-1-6654-1298-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/ISMAR-Adjunct54149.2021.00034

Multi-Drone Collaborative Trajectory Optimization for Large-Scale Aerial 3D Scanning

Fangping Chen[*]
Peking University

Yuheng Lu[†]
Peking University

Binbin Cai[‡]
Beijing Yunsheng Intelligent Technology Co., Ltd.

Xiaodong Xie[§]
Peking University

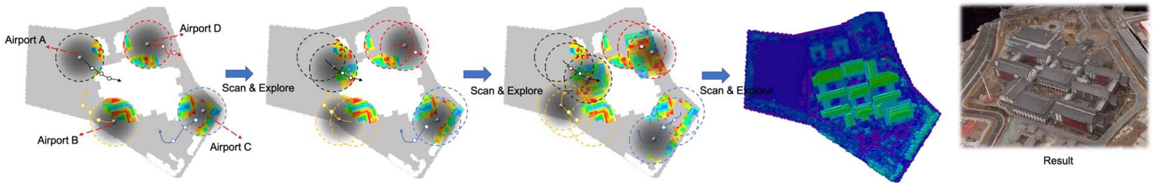


Figure 1: Four drones coordinated to scan and explore large-scale scenes, generate voxel maps, and independently plan the best viewpoints for three-dimensional reconstruction data shooting to form high-precision three-dimensional maps.

Abstract

Reconstruction and mapping of outdoor urban environment are critical to a large variety of applications, ranging from large-scale city-level 3D content creation for augmented and virtual reality to the digital twin construction of smart cities and automatic driving. The construction of large-scale city-level 3D model will become another important medium after images and videos. We propose an autonomous approach to reconstruct the voxel model of the scene in real-time, and estimate the best set of viewing angles according to the precision requirement. These task views are assigned to the drones based on Optimal Mass Transport (OMT) optimization. In this process, the multi-level pipelining in the chip design method is applied to accelerate the parallelism between exploration and data acquisition. Our method includes: (1) real-time perception and reconstruction of scene voxel model and obstacle avoidance; (2) determining the best observation and viewing angles of scene geometry through global and local optimization; (3) assigning the task views to the drones and planning path based on the OMT optimization, and iterating continuously according to new exploration results; (4) expediting exploration and data acquisition in parallel through multi-stage pipeline to improve efficiency. Our method can schedule routes for drones according to the scene and its optimal acquisition perspective in real-time, which avoids the model void and lack of accuracy caused by traditional aerial 3D scanning using routes of cultivating land regardless of the object, and lays a solid foundation for the 3D real-life model to directly become the available 3D data source for AR and VR. We evaluate the effectiveness of our method by collecting several groups of large-scale city-level data. Facts have proved that the accuracy and efficiency of reconstruction have been greatly improved.

Keywords: 3D reconstruction, Trajectory planning, Multi-drone coordination, the Best coverage viewpoint.

Index Terms: Computing methodologies—Modeling methodologies—;——Computing methodologies—Multi-agent planning—

1 Introduction

Recently, drones are increasingly used for large scale aerial 3D scanning and reconstructing. In order to obtain high-precision 3D reconstructions, a drone ought to take photos as densely as possible to exploit all available information about the scene. Manual piloting, even for a skilled human pilot, could not satisfy requirements on precision and viewpoints of each photo, while avoiding all kinds of obstacles in the environment. Additionally, the flight time of drones is limited so that pilots do not have enough time to think about how to shoot better and to adjust viewing angles.

Most of the software working for trajectory planning, which enables UAVs to collect data in a standardized manner, uses a lawnmower mode at a safe distance above the scene (i.e., buildings, cities). Under the traditional data-collecting method, the photos are taken autonomically with no awareness of the characteristics of the scene. This leads to over-sample some regions (i.e., rooftops and floors), while under-sampling others (i.e., facades, eaves, recesses, etc.), and therefore sacrifice reconstruction precision, and also results in model voids.

The scenes on surface are dynamic. If a drone lacks real-time perception and obstacle avoidance capabilities, relying only on historical coarse scene model for fine data collection, it is very likely that the drone hit obstacles [1, 2, 13, 25] as the increase of constructed structures and buildings on the surface. We propose a more time-sensitive data collection method to capture images to avoid feature matching errors in multi-view stereo algorithms because of light change. Meanwhile, considering the time limit in one drone's flight, we propose that multiple drones fly autonomously and work in parallel to achieve efficient data acquisition.

In light of the aforementioned observations, we come up with a whole new approach to deploy several automated drones in cities. When there is a need for modeling, multiple drones will take off from their automated airports in parallel to perform modeling tasks. Each drone can be used as a spherical scanner to perform real-time voxel reconstruction [4, 9] of surrounding environment in 360 degree. By setting different safety distances, spherical scanners can depict scene geometry with varying degrees of refinement. We use the classic, highest-quality transmission method [21] to subcontract drones, enabling them to collect data in parallel efficiently. The optimal camera positions and views for image capture based on the multi-view acquisition principle of MVS multi-view geometry and the resolution that needs to be modeled can be generated automatically in finely crafted scene geometry models. Through the traveling salesman principle [6], the system assigns the drone the shortest path channel to collect data. At the same time, we use the multi-stage flow

[*] e-mail: chenfangping@pku.edu.cn
[†] e-mail: yuhenglu@pku.edu.cn
[‡] e-mail: caibinbin@ikingtec.com
[§] e-mail: donxie@pku.edu.cn

978-1-6654-1298-8/21/$31.00 ©2021 IEEE
DOI 10.1109/ISMAR-Adjunct54149.2021.00034

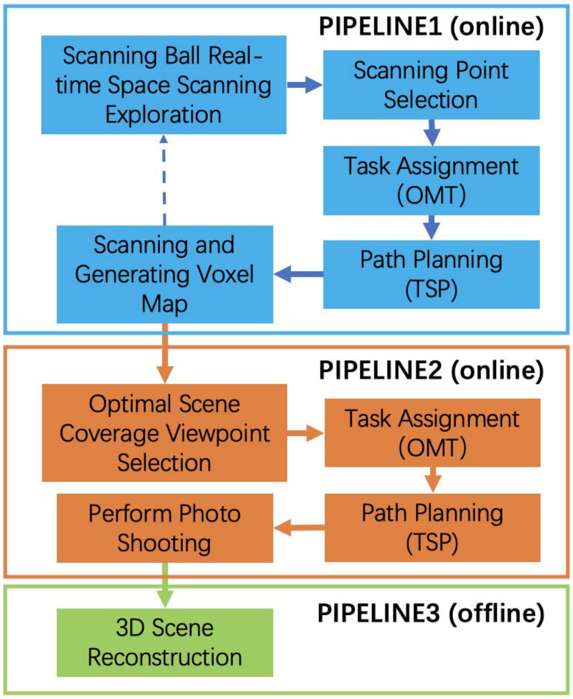


Figure 2: Algorithm flowchart: our method is divided into a three-stage pipeline, similar to the parallel design of a chip. The first stage of the pipeline is that the drone is used as a scanning ball to explore and generate voxel maps in real time. The second stage of the pipeline is to generate the optimal viewpoint based on the voxel map, and perform task allocation and optimal path optimization for multiple drones. The third stage of the pipeline is to reconstruct the real scene model in three dimensions.

structure in chip design [3] to trigger the modeling exploration and image capture of multiple drones in parallel, further improving the efficiency of data acquisition, and ensuring that the light of model reconstruction is as consistent as possible.

Two sets of experiments on both real and synthetic environments demonstrate the efficiency and accuracy of our method, and it can be easily extended to large-scale city-level high-precision mapping tasks. Furthermore, equipped with automatic airports, these large-scale maps can be updated regularly, which bridges the gap between the digital of the physical world automatically.

2 Related Works

Most of the aerial 3D reconstruction is based on the MVS multi-view stereo reconstruction algorithm, while the UAV simultaneous positioning and mapping algorithm is mostly used to avoid obstacles in the surrounding environment while navigating. Most of the methods do not combine the UAV's perception of the surrounding environment with the promotion of 3D reconstruction. Most of today's aerial 3D scanning modeling path planning software is not designed in conjunction with the geometric characteristics of the scene, but basically is in the lawnmower mode or the orbit mode [19, 23]. As a result, the modeling accuracy of most of the building facades and under the eaves is very poor.

Some 3D scanning path planning algorithms use orthophotos for segmentation to help UAVs perform modeling scan path exploration. The UAV intelligently flies above the eaves for fear of inaccurate estimation of the height of the building, which may cause the aircraft to hit the wall [16]. Some 3D scanning path planning algorithms use the shadow length of sunlight to determine the height of the building and then plan the flight path [27], but this situation subjects to environment limitations, once cloudy or haze occurs, the algorithm is invalid, and the estimated scene geometry is very rough, and it is difficult to qualitatively improve the modeling accuracy. The 3D scanning path planning method proposed by [22]works well, but it also relies on preliminary modeling for rough scene geometry estimation. Pre-modeling makes the whole process of the project unable to guarantee immediacy. After calculating the rough model, it proceeds further. The illumination of the scan may have already changed. In addition, the scene geometry generated by coarse modeling often ignores architectural details such as eaves and columns. And the efficiency of stand-alone operation is not high.

The classic multi-view stereo geometry [8, 10, 17, 18, 24, 26] will become the guiding ideology for the selection of the 3D reconstruction shooting path. Of course, there are some modeling strategies that use depth cameras [12] for real-time exploration, but the detection range of depth cameras is limited, and it also reflects infrared light. The sensitivity is therefore not good outdoors. The flight time of drones is limited. The next best view article [15] only considers the best viewing angle for photo collection, but ignores the algorithm of the shortest path connecting these best viewing angles, because the drone has a limited flight time. We can learn from the cooperative operation strategy of ground robots [5] to optimize our most efficient scanning strategy in three-dimensional space.

For this type of scenario where multiple entities and multiple tasks execute tasks concurrently, we refer to the multi-stage pipeline operation in the chip design [3] to improve operation efficiency. We combine local and global optimal strategies to coordinate optimization, not just using sub-model strategies [22].

3 Method

The existing technology basically treats the scene as known by default when scanning the scene. For example, as described in [19, 23], they treat the buildings in the city as a plane, and then scan the city by means of lawn mower, which will lead to too much data collected from the top of the building (oversampling), too little data collected from the facade and eaves of the building (missing sampling) or even missing, which will lead to holes in the reconstruction results. However, [22] mentioned that the circular trajectory was used to pre-scan the scene, which made it difficult to fully give feedback the geometric characteristics of the scene, and the optimal scene acquisition perspective was not accurate enough. The urban scenes reconstructed by the above methods can not measure most of the scenes quantitatively. The limited accuracy of the model and frequent holes in the model are the key factors that the current MVS model cannot be directly used as digital twin materials such as games, AR or VR.

In order to efficiently obtain the large-scale and high-precision 3D real scene model at the city level [11], our method is divided into the following three parts. The first part is to scan the surrounding environment into a voxel map by using the spatial spherical 3d scanner defined by us, which includes the construction of multi-purpose depth map scanning system, the allocation of multi-machine scanning tasks and the generation of the best path; In the second part, we will get the best shooting angle and path according to the scanned scene geometry; In the third part, we will cooperate with multiple unmanned aerial vehicles, optimize the multi-tasks in the first part and the second part, and further improve the efficiency of multi-machine collaborative data acquisition by adopting the method of chip design and multi-stage pipelining, As shown in Figure 2.

3.1 Spherical Space Scanner

We constructed a set of spherical scanning system as shown in Figure 3. The system consists of the drone and four pairs of binocular

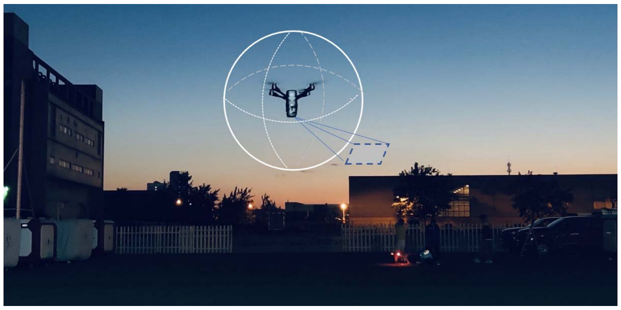


Figure 3: The drone-based spherical scanner sinks into the concave space for scanning

cameras on the facade, and also includes the modeling data acquisition camera at the bottom of the drone. The four pairs of binocular cameras consist of 8 4k wide-angle lenses with a 220-degree field of view. Combined with the algorithm, a sub-pixel-level precision depth map can be generated and mapped into the spherical space around the drone. With the help of TSDF truncation function algorithm (TSDF weight is a simple and effective measure of scanning quality accounting for both scanning distance and angle.), a three-dimensional voxel map of the surrounding environment is generated in real time. This is different from the traditional one-way exploration depth camera, but a spherical scanning volume is formed, which we will call the scanning ball for short here.

The scanning ball can not only generate voxel models in the surrounding 360-degree environment in real time, but also avoid obstacles in the surrounding environment, laying the foundation for safe and efficient path planning in the following part. Our algorithm uses the latest deep learning architecture [14] and is optimized on the NX hardware to achieve real-time performance. According to the designed baseline size, our best sensing range for scanning the ball is 50cm to 6m.

At the bottom of the scan ball is equipped with a Sony full-frame 60-megapixel pan/tilt camera (as shown in Figure 3). Collect the best photos of the best perspective of the surrounding environment. The focal length of the lens is 35mm. We can define the resolution of the captured photos based on this and the distance between the scanning ball and the surrounding geometric scene, so as to quantify the resolution of the captured data for each frame.

Here we use the judgement rules for the extraction of scanning ball scanning tasks in [5] to explore, and use the OMT method to subcontract multiple drones, that is, the tasks of multiple scanning balls, and then the large-scale NP-hard problem is reduced to a limited-scale single traveling salesman problem, as shown in the Figure 1.

3.2 The Best Coverage Viewpoint Model Generation

From the light field theory, when the camera faces the surface of the object as much as possible and shoots along the normal direction, the texture and surface features obtained are the best, which also conforms to the principle of three-dimensional reconstruction of multi-view geometry. While most of photos obtained in the modeling work relies on qualitative manual perception, and lacks quantitative standards, which results in uneven quality and low efficiency, our method features in quantitatively expressing the automatic generation of the best viewpoints.

First, taking the point cloud and voxel generated based on TSDF in the first part as input, we can adjust the size of the voxel block according to the requirements of modeling accuracy. And we can confirm the lateral overlap rate of the photo according to the camera's internal participating lens parameters. We can also set equal intervals for sampling, if the problem of how to quantitatively express the best shooting direction for the sampled surface points is solved.

The shooting direction of the drone's gimbal is generally defined as the yaw angle and the pitch angle. As shown in Figure 4(a), the blue curve x is the sampled surface point P on the surface of the object along the horizontal direction. The red curve y is the section line of the sampled point P on the surface of the object along the vertical direction. The best shooting direction is defined as solving the best yaw angle and pitch angle according to the object's surface characteristics.

Take the solution of the yaw angle in the horizontal direction as an example. As shown in Figure 4(b), the point cloud set of the horizontal section is denoted as B, and the centroid of B is denoted as A. The local section line (represented as point set T) and its tangent circle near the surface point P are shown in Figure 4(b), and O is the center of the circle.

I. Calculation of center coordinates and radius of tangent circle O

i. Calculate the normal direction for each point on the local section line T

$$T = neighbor(P, B, r) \tag{1}$$

ii. Each point in T is denoted as T_{P_i}, and the normal direction of each point is denoted as N_{P_i}

$$N_{P_i} = normal(T_{P_i}, T) \tag{2}$$

The normal function is:

Input: as shown in Figure 4(e), the surface point P and K points in the neighborhood. Fitting a straight line by the least square method:

$$l = argmin \sum_{i=1}^{K} d(T_{p_i}, l) \tag{3}$$

Output: the normal direction N is perpendicular to l and intersects with P.

iii. Calculate the intersection point between the normals of T_{P_i} and obtain the intersection set D

$$D = intersect(T_P, N_P) \tag{4}$$

The intersect function is:

Input: K normal straight lines n_1, n_2,...n_k

Output: the center of the circle, as shown in Figure 4(c)

$$O = mean(\sum_{i=1}^{K} \sum_{j=1}^{K} intersect(n_i, n_j)) \tag{5}$$

iv. Center $O = Mean(D)$, radius $r = OP$

II. Determine whether the circle tangent to the local section line is a circumscribed circle or an inscribed circle according to the three points: A, O, and P

i. When AO<AP, inscribed circle, as shown in Figure 4(b)

ii. When AO>AP, circumscribed circle, as shown in Figure 4(d)

III. Calculate the yaw angle of the gimbal

i. When the local section line and the circle are inscribed

$$yaw = \overrightarrow{PO} \tag{6}$$

$$Yaw = \theta = atan2(O_y - P_y, O_x - P_x) \tag{7}$$

ii. When the local section line is circumscribed to the circle

$$yaw = \overrightarrow{OP} \tag{8}$$

$$Yaw = \theta = atan2(P_y - O_y, P_x - O_x) \tag{9}$$

The same method can also be applied to the solution of to the vertical pitch angle.

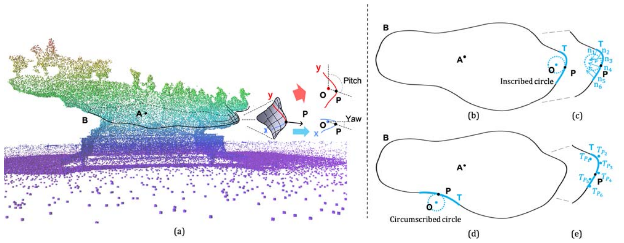


Figure 4: Generate the best viewpoint from the voxel model

3.3 Optimal coverage viewpoint allocation

After determining the mission viewpoint, we need to assign appropriate missions to multiple drones. This is a typical allocation problem. It is necessary to ensure that each viewpoint can be traversed one by one, but also to ensure that all drones have the highest overall execution efficiency and the lowest cost. Here we are based on the OMT problem and provide a good solution for minimizing the transport cost by considering the travel distance and endurance of the drone.

$$Cost_{min} = min \sum_{u=1}^{T} D(sp_u, vp_u) + \sum_{u=1}^{T} (E - e_u) * scale \tag{10}$$

where the $Cost_{min}$ is cost minimization, the $D(sp_u, vp_u)$ is the distance from the start point of the u-th drone to a set of view points, the e_u is the current endurance of the u-th drone, the *scale* is the scaling factor (here set *scale* = 2.4), the E is the normal endurance of the drone. By minimizing the cost function, we found the optimal allocation strategy to ensure the parallel operation efficiency and coordinated control of multiple drones.

After the drones have been assigned a set of viewpoints, we need to plan an optimal route for each drone based on the mission viewpoint. This route will traverse every mission viewpoint, and each viewpoint will only be traversed once. This is a typical TSP problem. The goal of TSP is to find a minimal cost path, it will visit each node and return to the starting node. Here we use the software package developed by [7], after a slight modification, to solve the TSP problem and find the optimal path.

4 Multi-drone Cooperative System

In order to fully schedule the collaborative modeling of multiple UAVs and improve work efficiency, we designed a clustered collaborative operation algorithm.

4.1 Three-stage pipeline architecture

The system architecture draws on the basic principles of parallel pipeline in chip design, and is divided into three-level pipeline according to the task volume and job characteristics.

The first level of pipeline: the scanning ball represented by the drone scans the area to obtain the voxel map of the scene. Based on these voxel maps, the camera's shooting viewpoint is re-planned. After the viewpoint tasks are clarified, according to the number of drones performing tasks in the second-level flowing water, each drone is reasonably allocated its own viewpoint tasks.

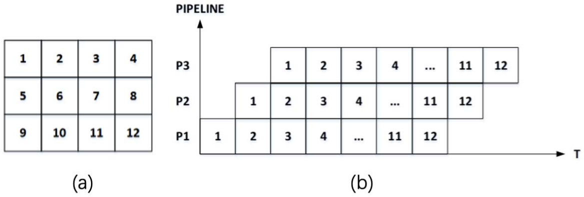
The second level of pipeline: the task is issued to each drone, and each drone reasonably plans its own execution route for the viewpoint task to be performed by itself, and completes the scanning and shooting tasks in this area efficiently and in parallel.

The third level of pipeline: the drone sends back the photos taken to the cloud server in real time, and allocates the corresponding nodes according to the computing power of the server and the number of photos to start the modeling task.

Three-stage pipeline, as shown in Figure 2.

4.2 Execution process and parallel time

First, divide the area to be modeled according to the terrain. The area shown in the Figure 5(a) is divided into 12 small blocks to ensure that the drone's endurance time is greater than the operation time of one area.



1	2	3	4
5	6	7	8
9	10	11	12

P3			1	2	3	4	...	11	12
P2		1	2	3	4	...	11	12	
P1	1	2	3	4	...	11	12		

Figure 5: (a) Schematic diagram of modeling block division; (b) Schematic diagram of cluster flow operation.

Then the drones start clustering parallel pipeline operations. As shown in Figure 5(b), the task of one drone responsible for the first stage of pipeline is Pipeline 1 (P1), and the task of multiple drones responsible for the second stage of pipeline is Pipeline 2 (P2), the task of real-time image transmission and 3D reconstruction of the server is Pipeline 3 (P3). When P1 of area 1 is executed, P2 of area 1 starts to execute, and P1 of area 2 starts to execute at the same time. When P2 of area 1 is executed, P3 of area 1 starts to execute, and P2 of area 2 starts to execute at the same time.

5 Experiment

We conducted 3 sets of experiments in the real and synthetic environment respectively. The endurance of our drone is 50 minutes, and we set the speed to 5 m/s. That means the longest route for each flight is 15km. Each drone is equipped with an airport, which can automatically replace the battery of the drone without any human intervention in the whole procedure. Compared with the real environment, the synthetic environment can provide accurate environmental coordinates, which can be used to quantify the error of the reconstruction model.

5.1 Experiment in the real environment

Buildings with eaves and sculptures with complex structures are the typical failure cases of traditional path planning method. We selected a pavilion with eaves and a complicated sculpture for the first experiment of the real environment. Compared with the traditional flight path (as shown in Figure 6(a)), our method (Figure 6(b)) builds the voxel map of the pavilion, and then plans the shooting path according to local structure. As shown in Figure 6(c)(d), the reconstructed model of our method is more clear, especially under the eaves. Similarly, Figure 7, shows the comparison of the reconstructed model of the complicated sculpture, the model geometry of our method is richer.

For the multi-drone cooperative system, we select a large-scale classical building complex (Hilton Hotel) for the second experiment of the real environment. The building complex covers an area of 140,000 square meters and has a circumference of 1,561.385 meters. We divided it into 12 blocks. 4 drones, numbered A, B, C and D respectively, take off from the airport automatically. They first explore the local surrounding environment, generating a voxel map, as shown in Figure 1. The optimal data acquisition viewpoints are generated based on the map. Then the multi-drone collaborative pipeline is started, finishing the data acquisition of all blocks one by one. During the data acquisition, we found that the drone could descend into the courtyard to collect the image under the eaves.

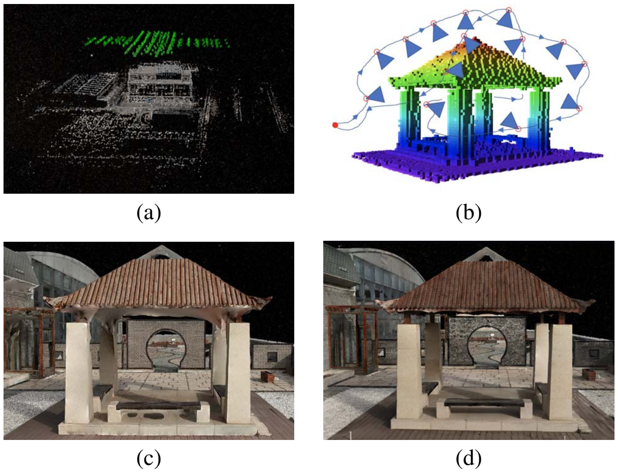


Figure 6: Two different 3D reconstruction trajectories and reconstruction results

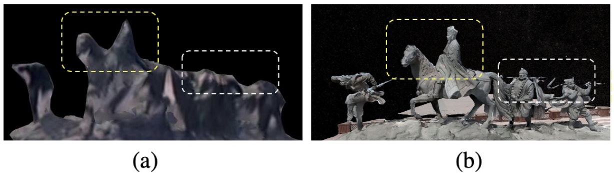


Figure 7: Details of the different reconstruction results of the two methods

Figure 8, shows the reconstruction model of eaves under different path planning strategies.

5.2 Experiment in the synthetic environment

Compared with the real environment, the synthetic environment can provide accurate environmental coordinates, which can be used to quantify the error of the reconstruction model.

We conduct experiments in the simulation scene of a video game emulator. In the simulation scene, we can acquire the ground truth geometry and appearance easily. By comparing the reconstructed model with the ground truth, we are able to obtain the quantitative comparison result. We selected a classical Chinese building as the experiment target, using the Unreal Engine [7] as the Emulator and using UnrealCV [20] Python Library as the image rendering tool. Like the previous experiment configuration, we chose the lawnmower strategy as the baseline. For the collected images, we used the same method as the real environment experiment to generate mesh model and point cloud, as shown in Figure 9(a), (b). In order to compare the difference between reconstruction model and Ground Truth, we subtracted Ground Truth from modeling point cloud. To be specific, we first sampled Ground Truth mesh into point cloud through Monte Carlo Sampling. Then, the voxel filter was used to control the point cloud density (the size of voxel was 0.1), and we only kept 1 point in each voxel. After that, we used feature points to roughly register the ground truth and the reconstructed point cloud, which were further finely registered using the iterative nearest point (ICP) algorithm. Finally, for each point in the reconstructed point cloud, we searched for the nearest point in the ground truth point cloud, and counted the distribution of the distance to the nearest point, which was regarded as subtraction result. As shown in Figure 9(c), red points represent the large distance, and the blue represent the small. Figure 10(b) shows the distance distribution of the Lawnmower strategy and ours. The experimental results show that our method is closer to Ground Truth, especially in occluded areas, such as areas under the eaves.

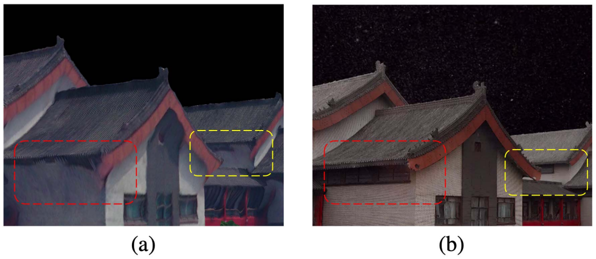


Figure 8: Comparison of modeling results between traditional method(a) and our method(b)

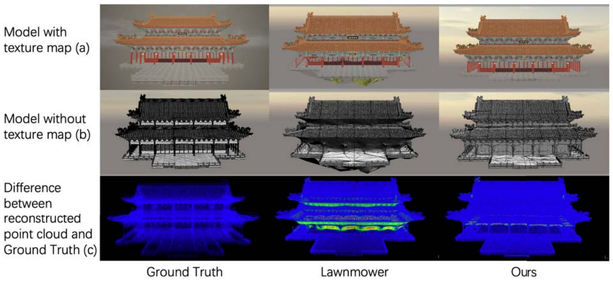

Model with texture map (a)
Model without texture map (b)
Difference between reconstructed point cloud and Ground Truth (c)
Ground Truth
Lawnmower
Ours

Figure 9: Visualization of simulation experiment

5.3 Multi-stage pipeline modeling time optimization

The execution time of the first stage pipeline is 18 min, and the execution time of the third stage pipeline is 21 min. Set the number of drones (N) according to the number of viewpoints and the execution time of the other two stages of pipeline. The execution time is basically the same to ensure the parallel flow of drones.

The calculation execution time is:

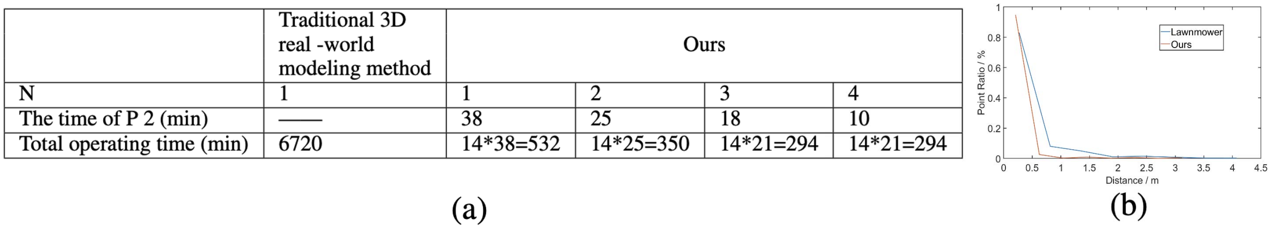
$$T_{total} = (B_{num} + P_{num} - 1) * T_{p\_max} \tag{11}$$

where B_{num} is the number of blocks, P_{num} is the number of pipeline stages, T_{p_max} is the maximum execution time of pipeline stages. Here $P_{num} = 3$.

As shown in Figure 10(a), when N=1, the operating time is basically the same as that of traditional oblique photography; when N=2, 3, and 4, the operating efficiency is the highest, which is 63% and 53% of the original time respectively. When $N \geq 3$, the efficiency is no longer improved because it is limited by the execution time of the other two pipelines. In the choice of N, the system will automatically make a flexible choice based on the size of the block, the number of viewpoints and the other two-stage pipeline time.

6 Conclusion

We propose a method for multi-drone coordination and large-scale three-dimensional spatial refinement and high-efficiency scanning modeling. This method can not only perceive and finely reconstruct the geometric characteristics of the scene in real time, but also generate the best coverage viewpoints based on these characteristics. Multiple drones work together according to the optimal quality transmission theory, and adopt a multi-stage pipeline design method.



	Traditional 3D real -world modeling method	Ours			
N	1	1	2	3	4
The time of P 2 (min)	——	38	25	18	10
Total operating time (min)	6720	14*38=532	14*25=350	14*21=294	14*21=294

(a)

(b)

Figure 10: (a) Experimental result; (b) The distance distribution of the Lawnmower strategy and ours.

Drones can perform tasks concurrently, achieve high-efficiency and high-precision modeling and scanning of the largest-scale scenes, and eventually create a high-precision medium for the digitalization of the physical world. By scanning multiple sets of real large-scale urban scenes for experimental testing, it has proved that our method is far superior to traditional methods in reconstruction quality. Our method not only improves the quality of geometric model, but also optimizes the visualization effect of model texture, and it can also effectively reduce MVS algorithm errors caused by illumination changes.

In the near future, we believe that the data collection cluster composed of drones and airports, which can cooperate with each other as efficiently as small satellites, can automatically refresh the real 3D data of the Earth's surface regularly with high precision and high frequency, so that we can create more unexpected value on this digital four-dimensional medium.

Acknowledgments

This work is partially supported by the Beijing Major Science and Technology Project under contract No. Z191100010618003 and the Project 2019BD004 supported by PKU-Baidu Fund.

References

[1] A. J. Barry, H. Oleynikova, D. Honegger, M. Pollefeys, and R. Tedrake. Fast onboard stereo vision for uavs. In *Vision-based Control and Navigation of Small Lightweight UAV Workshop, International Conference On Intelligent Robots and Systems (IROS)*, 2015.

[2] A. Bhoi. Monocular depth estimation: A survey. *arXiv preprint arXiv:1901.09402*, 2019.

[3] V. A. Chandrasetty. *VLSI design: a practical guide for FPGA and ASIC implementations*. Springer Science & Business Media, 2011.

[4] M. Dharmadhikari, T. Dang, L. Solanka, J. Loje, H. Nguyen, N. Khedekar, and K. Alexis. Motion primitives-based path planning for fast and agile exploration using aerial robots. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 179–185. IEEE, 2020.

[5] S. Dong, K. Xu, Q. Zhou, A. Tagliasacchi, S. Xin, M. Nießner, and B. Chen. Multi-robot collaborative dense scene reconstruction. *ACM Transactions on Graphics (TOG)*, 38(4):1–16, 2019.

[6] M. Dorigo and L. M. Gambardella. Ant colonies for the travelling salesman problem. *biosystems*, 43(2):73–81, 1997.

[7] E. G. U. Engine. http://www.unrealengine.com.

[8] Y. Furukawa and C. Hernández. Multi-view stereo: A tutorial. *Foundations and Trends® in Computer Graphics and Vision*, 9(1-2):1–148, 2015.

[9] L. Heng, A. Gotovos, A. Krause, and M. Pollefeys. Efficient visual exploration and coverage with a micro aerial vehicle in unknown environments. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1071–1078. IEEE, 2015.

[10] B. Hepp, M. Nießner, and O. Hilliges. Plan3d: Viewpoint and trajectory optimization for aerial multi-view stereo reconstruction. *ACM Transactions on Graphics (TOG)*, 38(1):1–17, 2018.

[11] R. Huang, D. Zou, R. Vaughan, and P. Tan. Active image-based modeling with a toy drone. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6124–6131. IEEE, 2018.

[12] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison, et al. Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pp. 559–568, 2011.

[13] E. Kaufmann, A. Loquercio, R. Ranftl, A. Dosovitskiy, V. Koltun, and D. Scaramuzza. Deep drone racing: Learning agile flight in dynamic environments. In *Conference on Robot Learning*, pp. 133–145. PMLR, 2018.

[14] A. Kendall, H. Martirosyan, S. Dasgupta, P. Henry, R. Kennedy, A. Bachrach, and A. Bry. End-to-end learning of geometry and context for deep stereo regression. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 66–75, 2017.

[15] M. Krainin, B. Curless, and D. Fox. Autonomous generation of complete 3d object models using next best view manipulation planning. In *2011 IEEE International Conference on Robotics and Automation*, pp. 5031–5037. IEEE, 2011.

[16] Q. Kuang, J. Wu, J. Pan, and B. Zhou. Real-time uav path planning for autonomous urban scene reconstruction. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1156–1162. IEEE, 2020.

[17] N. C. Mithun, K. Sikka, H.-P. Chiu, S. Samarasekera, and R. Kumar. Rgb2lidar: Towards solving large-scale cross-modal visual localization. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 934–954, 2020.

[18] R. Mur-Artal and J. D. Tardós. Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. *IEEE Transactions on Robotics*, 33(5):1255–1262, 2017.

[19] Pix4Dcapture. http://pix4d.com/product/.

[20] W. Qiu and A. Yuille. Unrealcv: Connecting computer vision to unreal engine. *European Conference on Computer Vision*, 2016.

[21] S. T. Rachev and L. Rüschendorf. *Mass Transportation Problems: Volume I: Theory*, vol. 1. Springer Science & Business Media, 1998.

[22] 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*, pp. 5324–5333, 2017.

[23] D. Robotics. http://3dr.com.

[24] J. L. Schonberger and J.-M. Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4104–4113, 2016.

[25] N. Smith, N. Moehrle, M. Goesele, and W. Heidrich. Aerial path planning for urban scene reconstruction: a continuous optimization method and benchmark. 2018.

[26] C. Strecha, W. Von Hansen, L. Van Gool, P. Fua, and U. Thoennessen. On benchmarking camera calibration and multi-view stereo for high resolution imagery. In *2008 IEEE conference on computer vision and pattern recognition*, pp. 1–8. Ieee, 2008.

[27] X. Zhou, K. Xie, K. Huang, Y. Liu, Y. Zhou, M. Gong, and H. Huang. Offsite aerial path planning for efficient urban scene reconstruction. *ACM Transactions on Graphics (TOG)*, 39(6):1–16, 2020.