

# Capturing, Reconstructing, and Simulating: the UrbanScene3D Dataset

Liqiang Lin<sup>✉</sup>, Yilin Liu<sup>✉</sup>, Yue Hu, Xingguang Yan, Ke Xie, and Hui Huang<sup>\*✉</sup>

Shenzhen University

**Abstract.** We present *UrbanScene3D*, a large-scale data platform for research of urban scene perception and reconstruction. *UrbanScene3D* contains over 128k high-resolution images covering 16 scenes including large-scale real urban regions and synthetic cities with  $136 \text{ km}^2$  area in total. The dataset also contains high-precision LiDAR scans and hundreds of image sets with different observation patterns, which provide a comprehensive benchmark to design and evaluate aerial path planning and 3D reconstruction algorithms. In addition, the dataset, which is built on Unreal Engine and Airsim simulator together with the manually annotated unique instance label for each building in the dataset, enables the generation of all kinds of data, e.g., 2D depth maps, 2D/3D bounding boxes, and 3D point cloud/mesh segmentations, etc. The simulator with physical engine and lighting system not only produce variety of data but also enable users to simulate cars or drones in the proposed urban environment for future research. The dataset with aerial path planning and 3D reconstruction benchmark is available at: <https://vcc.tech/UrbanScene3D>

**Keywords:** UAV; urban scene dataset; aerial path planning; 3D acquisition; 3D reconstruction; city simulation

## 1 Introduction

With the development of digital photography and 3D scanning technologies, we have witnessed the explosive growth of data in recent years. Rich data sources and interaction methods bring rapid research progress in computer vision, computer graphics and robotics. For indoor scenes, with the help of sufficient data and real-time interactions [40,1,30,28], many fundamental problems, such as 2D/3D object detection and segmentation [16,33], depth estimation [24,43], 3D reconstruction [9,14,8,7,41] and autonomous navigation [50,6,15], have been better solved in a data-driven manner. However, things are different for outdoor study. The lack of effective devices and the extensive scales dramatically increase the difficulty of outdoor data capturing [48]. Also, due to the varied weather and light conditions, the outdoor scenes change fast and thus pose additional challenges to structure the data and design robust acquisition algorithms.

The current outdoor datasets are usually built by onboard equipment [13,10,38], e.g., RGB cameras and/or LiDAR. Nonetheless, it is challenging to use these

---

<sup>\*</sup> Corresponding author



Fig. 1: A glance of synthetic (top) and real (bottom) scenes in UrbanScene3D.

sensors to thoroughly capture the whole environment due to the limited field of views and routing choices. Other urban datasets [48,37] constructed by 3D modelers are usually clean and complete, but the models inside generally lack geometric and textural details. The gap between the proposed datasets and real-world scenarios remains large.

To tackle these problems, we present a large-scale urban scene dataset, *UrbanScene3D*, which consists of both man-made and real-world reconstruction scenes in different scales, together with a convenient simulator built on Unreal Engine and AirSim. The man-made scene models have compact structures, which are carefully constructed/designed by professional artists; see the top of Fig. 1. Probably more important, *UrbanScene3D* also offers dense, detailed scene models reconstructed by aerial images through multi-view stereo (MVS) techniques; check out the bottom of Fig. 1, where the scene models are with realistic textures and meticulous geometric structures.

In particular, to investigate how to better acquire and reconstruct outdoor scenes, we select a set of scene representatives and capture them with a drone flying along a set of aerial paths. These flights are calculated by different planning algorithms for 3D urban scene reconstruction. Thus, for each representative environment, we are able to provide a variety of reconstructed meshes with the corresponding scene observations (aerial acquisition paths and the captured image sets). Besides, with the help of a high-precision laser scanner applied in the real world and the synthetic ground-truth models, we have constructed a benchmark that provides the point-level accuracy and completeness analysis of each reconstructed mesh. This enables a robust evaluation of both path planning

strategies and MVS algorithms. In addition, the physical engine of AirSim enables users to simulate the robots (cars/drones) and test a variety of autonomous tasks in the proposed environments. The involvement of both synthetic and real scenes effectively extends the generalization ability of resulting algorithms.

In summary, our main contributions include: i) a large-scale urban scene dataset (Sec. 3) that facilitates research in various areas; ii) a comprehensive benchmark (Sec. 5) for investigating the impact of different factors in aerial path planning (Sec. 4) for 3D urban reconstruction; iii) an easy-to-use simulation platform (Sec. 6) for autonomous driving, robotics, and embodied AI research.

## 2 Related Work

*Outdoor datasets.* The fast development of autonomous driving involves enormous outdoor datasets [38,13,39,2,10]. They often offer stereo sequences, 3D LiDAR point clouds, camera calibration, and 3D object tracklet labels for outdoor scenes, thus promoting many applications and research. Although these ground-based sensors can capture small-scale scenes [19], they usually have very restricted views and routing choices, which cause significant challenges to cover large-scale urban areas. Meanwhile, the unmanned aerial vehicles (UAVs) or drones, have much better visibility and more freedom to maneuver, making them suitable for a complete coverage of wide regions. However, most existing UAV based datasets are not designed for capturing the entire scene. Instead, they only provide partial observations for perception tasks, such as semantic segmentation [25], object detection [12,49,26], action recognition [5], gesture recognition [32,31], or tracking [3,29]. On the contrary, we carefully plan the drone path to obtain a complete capture, from which the entire 3D scene can be reconstructed with MVS methods.

*Synthetic CAD datasets.* Different from real-world datasets, building synthetic datasets with CAD models [21,4] can offer a complete structured environment at a much lower cost, but lack geometric and textural details. To bridge the gap, HoliCity [48] aligns the real-world panoramas with the CAD models to provide real texture. However, the discrepancy is still significant since the geometry is too coarse and the panoramas only covers a portion of the entire scene. Instead, our UrbanScene3D contains both real-world and synthetic CAD scenes, facilitating research for both high-quality urban reconstruction and holistic scene understanding. Similar to Mueller et al. [29], we also provide a simulator that stimulates the research of online real-time capturing and understanding of 3D urban scenes.

*Aerial path planning for urban scene capture.* To capture urban scenes with drones, aerial path planning plays a vital role; see a recent survey [47]. Manual control and Zig-Zag patterns are inefficient, difficult to achieve decent coverage, and challenging to fulfill practical factors, such as safety restriction and battery capacity. To deal with these problems, existing methods [36,34,17,37,20,46,44]

Table 1: Statistics of UrbanScene3D with 10 synthetic and 6 real scenes. Area: the covered area of the scene model; Size: the size of the scene model; Texture: the number of texture images contained in the scene model; Texture: the size of texture; Object: the number of objects in the scene model.

Scene	Area( $m^2$ )	Size(Mb)	Texture(#)	Texture(Mb)	Object(#)
New York	$7.4 \times 10^6$	86	762	122	744
Chicago	$2.4 \times 10^7$	146	2277	227	1629
San Francisco	$5.5 \times 10^7$	225	2865	322	2801
Shenzhen	$3.0 \times 10^6$	50	199	73	1126
Suzhou	$7.0 \times 10^6$	191	395	24	168
Shanghai	$3.7 \times 10^7$	308	2285	220	6850
School	$1.7 \times 10^4$	25	47	488	3
Bridge	$1.3 \times 10^4$	28	237	44	8
Castle	$7.0 \times 10^3$	9	47	184	6
Town	$4.4 \times 10^3$	112	73	348	17
Campus	$1.3 \times 10^6$	1859	122	3676	178
Residence	$1.0 \times 10^5$	356	52	1760	34
Square	$1.1 \times 10^6$	3665	799	980	156
Hospital	$5.0 \times 10^5$	6266	94	744	114
Polytech	$1.5 \times 10^4$	3523	50	221	3
ArtSci	$1.6 \times 10^4$	25395	118	556	3

optimize the drone path according to certain goals and constraints with a coarse proxy model or a top view image as input. Specifically, Smith et al. [37] designs an optimization objective for better multi-view stereo results, ensuring the completeness and accuracy of the reconstruction. Further, Zhang el al. [44] propose a continuous path planner to adequately shorten the path length and reduce sharp turns. To be able to do offsite planning, Zhou et al. [46] utilize a satellite image to estimate a 2.5D proxy for view selection. For online planning, the researchers learn to construct 2.5D height maps [23] or estimate 3D bounding boxes of buildings [22] on-the-fly for drone navigation and exploration. Due to the lack of valid data, these methods usually rely on a handcraft heuristic function to optimize the capturing views or the flight trajectories. However, the quality of the final reconstruction is constrained by the high-order relation among observations, which is difficult to model by heuristics and optimization designs. Our data and benchmark will efficaciously facilitate future research on this topic.

### 3 The UrbanScene3D Dataset

The goal of UrbanScene3D is to provide a general data platform for 3D vision, graphics and robotics research in urban scene environments with different scales. UrbanScene3D provides 10 synthetic and 6 real-world scenes with CAD and reconstructed mesh models and the corresponding aerial images; see Fig. 1 and

Table 2: More path planning statistics of 4 synthetic and 2 real representative scenes selected from UrbanScene3D. Tri: the number of triangles of the scene model; Proxy: the number of different levels of proxies provided; Overlap: the overlap rate used to sample the proxies; Planner: the number of different planners used to generate aerial paths; Path: the total number of generated flight paths of the scene; Image: the total number of captured images of the scene.

Scene	Tri(#)	Proxy(#)	Overlap(%)	Planner(#)	Path(#)	Image(#)
School	250,282	4	70 & 90	4	25	14,897
Bridge	394,022	4	70 & 90	4	25	13,228
Castle	109,513	4	70 & 90	4	25	7,414
Town	1,197,751	4	70 & 90	4	25	8,948
Polytech	2,613,608	3	90	4	13	19,635
ArtSci	4,524,028	3	90	4	13	12,261

Table 1 for more details. The synthetic CAD scenes consist of various compact primitive structures including buildings, bridges, streets, and vegetation, all of which are built by professional artists. For real scenes, we use a drone to capture images for MVS. We program the drone to follow an aerial path generated with oblique photography, a commonly used industrial solution conducted by *DJI-Terra*<sup>1</sup> for 3D city acquisition. Based on these images, we reconstruct the scenes with *ContextCapture*<sup>2</sup>, a commercial MVS solution. Specially, the selected representative scenes for our benchmark include additional huge volume of capture data using various path planners under different settings; see Table 2.

For oblique photography planner, we use a professional *DJI M300RTK*<sup>3</sup> drone loaded with five HD *PSDK 102S* aerial cameras. For the other planners, we use a single-camera *DJI PHANTOM 4 RTK*<sup>4</sup> drone.

Table 3 summarizes the difference between UrbanScene3D and the existing outdoor datasets. We further highlight the features of UrbanScene3D as follows.

*Up in the air.* Most outdoor datasets are ground based and the capture usually do not cover the entire scene. The capture is incomplete even for aerial dataset like UAVDet [12], since its goal is object detection. Similar to BlendedMVS [42], we provide aerial captures that are suitable for MVS. Specifically for the representative scenes, we adopt multiple optimized aerial paths for the capture, which lead to greater quality urban scenes. Our approach not only provides a high-quality dataset for local perception research but also enables the research of global understanding for real 3D urban scenes.

<sup>1</sup> <https://www.dji.com/dji-terra>

<sup>2</sup> <https://www.bentley.com/en/products/brands/contextcapture>

<sup>3</sup> <https://www.dji.com/matrice-300>

<sup>4</sup> <https://www.dji.com/phantom-4-rtk>

Table 3: Comparing UrbanScene3D with existing 3D outdoor datasets. Area stands for the maximum area across all scenes in the dataset. Path stands for the number of flights/rides for capturing. Note that for BlendedMVS, we only show the estimated statistics of the urban scenes.

Datasets	Scene	Type	Area( $km^2$ )	Path(#)	Image(#)	Diversity	Simulator	Semantics
KITTI [13]	driving	LiDAR	/	1	93k	5 scans	/	/
Cityscape [10]	driving	stereo	/	/	25k	50 cities	/	2D
HoliCity [48]	urban	CAD	20	/	6.3k	1 city	/	3D
BlendedMVS [42]	mixed	MVS	0.02	29	17k	29 scenes	/	/
SYNTHIA [35]	driving	CAD	/	/	50k	1 city	car	3D
Mueller2016 [29]	mixed	CAD	/	/	/	/	drone	3D
Smith2018 [37]	urban	CAD	0.03	/	/	5 scenes	/	/
<b>UrbanScene3D</b>	urban	CAD&MVS	55	130	128k	16 cities/scenes	car&drone	3D

*Extensive scale.* Many existing datasets do not offer the complete capture of large-scale scenes; see Table 3. While Holicity offers a  $20km^2$  scene of London, our dataset includes three large-scale city scenes that cover above  $24km^2$  area. Meanwhile, we offer multiple real-world complete scenes, which include two extensive scenes that cover above  $1km^2$  area, whose scale is unseen in previous datasets like BlendedMVS [42]. More diverse urban scenes and the corresponding aerial images with shot poses are also available. Additionally, our drone flight path length can be up to  $17km$ , which benefits the research for SLAM or SfM.

*Path planning research.* For current path planning methods, factors such as the proxy accuracy and overlap rate play an vital role. Our path planning benchmark specifically include different settings for these factors. We show their influence on the final reconstruction quality and acquisition efficiency; see Sec. 5. In contrast, existing benchmark [37] for path planning does not consider these factors and only includes synthetic scenes. Besides, our benchmark includes hundreds of flight paths, which will boost the future research for path planning techniques.

*Multiple captures.* The different weather and lighting conditions in outdoor environment pose huge challenges for perception and reconstruction in general. Our benchmark offers multiple drone flights for each scene over different times, this greatly increases the variety of data, which can benefit learning based methods to in turn solve this problem. For example, the abundance (10k+) of the real scene images in the benchmark under different lighting conditions allows the future research for NeRFs [27] that can decouple geometry, material and light.

*Simulation environment.* The sim-to-real gap is a critical problem for the robotics and embodied AI research. Our simulator can directly import our real-world scenes and simulate the drones inside them, hence this gap would be greatly shortened. In contrast, most existing simulator for autonomous driving or UAVs operate only for virtual scenes. Further, our simulator can show the coverage of the scene in real-time, which is useful for research on UAV exploration.

## 4 Scene Acquisition with Aerial Path Planning

UrbanScene3D offers a wide variety of potential applications, from instance segmentation, multi-view stereo, to depth estimation and novel view synthesis. See Sec. 6 for further details. In this section, we extend the UrbanScene3D dataset to a benchmark to evaluate the quality and efficiency of different path planning algorithms, as well as the impact of input proxies in Sec. 5.

UrbanScene3D contains all the data needed for testing path planning algorithms and analyzing the reconstructed results, including proxies, ground-truth models/point clouds, paths, images, and reconstructed results; see Table 2.

The majority of path planners usually rely on a pre-computed coarse model (also called *proxy* [37]) of the environment. The proxy can be obtained by various methods, including a quick reconstruction after a simplified oblique photography pass [37,34], satellite image [46], map providers or even through a real-time reconstruction [22]. The quality of the proxy, including the accuracy of the topology and the face normal, can affect the final quality of the reconstruction.

In Sec. 4.1, we briefly introduce the 4 path planners we used to generate different trajectories. To evaluate the influence of the proxy in the path planning, we then introduce different proxies we used in the dataset in Sec. 4.2.

### 4.1 Aerial Path Planning Methods

In order to offer more paths and images, we use four different path planners, including oblique photography and the methods proposed by Smith et al. [37], Zhou et al. [46], and Zhang et al. [44], to generate the paths on the different proxies mentioned in Sec. 4.2 with different overlap rates for different scenes, resulting in 100 different paths for synthetic scenes and 26 paths for real scenes. Fig. 2 shows an illustration of the paths generated with the four different planners on the synthetic scene School and the real scene Polytech.

*Oblique photography.* Given the image overlap and ground sample distance (GSD), Oblique photography generate an S-shaped trajectory at a fixed height (calculated by the GSD) and compute the required capture location. The S-shaped trajectory is usually computed by *Complete Coverage Path Planning* (CCPP) algorithm, which could guarantee the complete coverage of an area even with an irregular shape [45]. Also, the number of turns is minimized, which could significantly increase the capturing efficiency.

*Planner proposed by Smith et al. [37].* Unlike Oblique photography, Smith et al. [37] directly optimize viewpoints according to a heuristic function, *reconstructability*, which consider both the potential error of triangulation and feature matching in the *Multi-view Stereo* (MVS) process. In each iteration, Smith et al. [37] first compute the so-called *reconstructability* of each point respect to the current viewpoint set. Then they try to maximize this measurement by adjusting the position and orientation of each viewpoint.

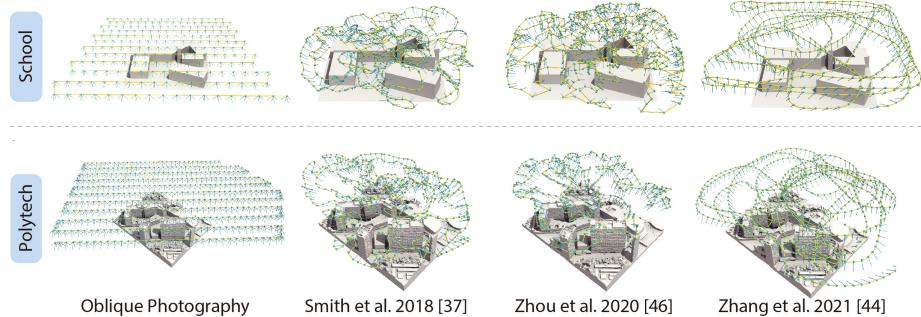


Fig. 2: The comparison of the paths generated with different aerial planning methods on the synthetic scene School and the real scene Polytech.

*Planner proposed by Zhou et al. [46].* Similar to Smith et al. [37], Zhou et al. [46] also consider the reconstructability of each point during the planning. However, they only use this measurement to reduce useless viewpoints. They first generate a large viewpoint set, which suppose to be *complete* but highly *redundant*. In each iteration, they define the *view redundancy* on the computed reconstructability and delete the most redundant viewpoint accordingly.

*Planner proposed by Zhang et al. [44].* Compared with Oblique photography, Smith et al. [37] and Zhou et al. [46] generate trajectories with higher capturing quality. However, the heuristic function which is defined on the viewpoints brings many sharp turns in the final trajectories. The drastic speed change significantly decreases the capturing efficiency. Thus, Zhang et al. [44] involves path smoothness in the heuristic function and utilize *Rapidly-exploring random* (RRT) tree to search for an efficient and high-quality trajectory.

#### 4.2 Geometric Proxies

The proxies are essential for aerial path planning methods. Detailed proxy usually leads to much better reconstruction results. Previous works either use a rough scene proxy [44,37] or a 2.5D model extracted from satellite images [46] to plan the path. UrbanScene3D provides proxies in different levels of details, which could be used to analyze the relations between proxies, paths, and the corresponding qualities of the reconstructed models; see Fig. 3.

The box proxy (*box*) is the roughest level of proxy with incorrect topology. It is built by replacing the building set in the scene with its bounding box. For the real scene, the box proxy is deprecated due to the safety issue. Similar to *box* proxy, *coarse* proxy is also built by finding the bounding box of each building in the scene. However, the *coarse* proxy has more accurate topology, which might lead more accurate path planning result. The intermediate proxy (*inter*) is reconstructed by downsampled images which are captured by oblique photography. The ground-truth meshes (*fine*) for the synthetic scenes can also be

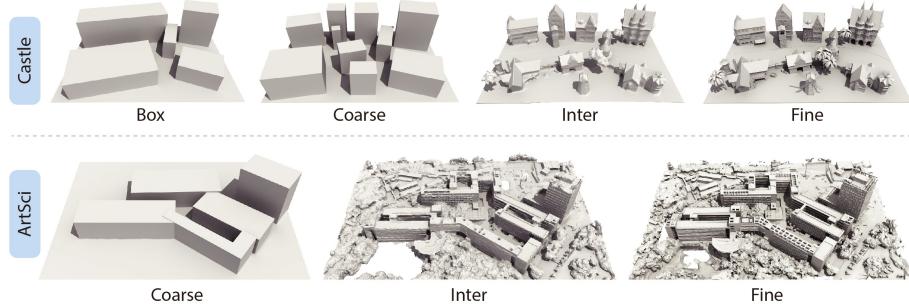


Fig. 3: The proxies of scene School in different levels of details. Box: the roughest proxy; Coarse: the coarse level of proxy; Inter: the intermediate level of proxy; Fine: the finest level of proxy.

used as proxies, which is supposed to provide the largest geometric information during the path planning process. The *fine* proxy for the real scenes are reconstructed with non subsampled images captured by oblique photography. Since there are no ground-truth meshes for real scenes, we use reconstructed models by high-density images as their corresponding fine proxies.

## 5 Scene Reconstruction Benchmarks

In this section, we evaluate the paths generated with different path planning methods for both energy cost, aerotriangulation accuracy, and reconstruction quality, providing the point-level accuracy and completeness analysis of the reconstructed mesh. The statistics of energy consumption of UAV capturing is given in Sec. 5.2. In Sec. 5.3, we analyze the aerotriangulation results of different planners. And finally, we evaluate the reconstruction results in Sec. 5.4. The Sec. 5.5 give a overall comparison of all the four planners.

Other information, e.g., the evaluation of different overlaps, the cost of model reconstruction, and the reconstruction evaluation on the other scenes are included in the supplementary material.

### 5.1 High-precision LiDAR Scan

The synthetic scenes come with ground-truth meshes for evaluation. For the real scenes, we scan the entire building with high-precision LiDAR scanners loaded with GPS localization devices. The point clouds are then registered with each other, resulting in a high-precision LiDAR scan of the whole building.

The LiDAR scanner is Trimble X7 with self-calibration and self-registration techniques. The ranging noise is  $0.5\text{mm}$ , the ranging accuracy is  $2\text{mm}$ , the angular accuracy is  $21''$ , and the accuracy of 3D points is  $1.5\text{mm}$  at  $10\text{m}$  and  $2.4\text{mm}$  at  $20\text{m}$ . Each scan, including self-calibration, takes 2 minutes and 34 seconds.

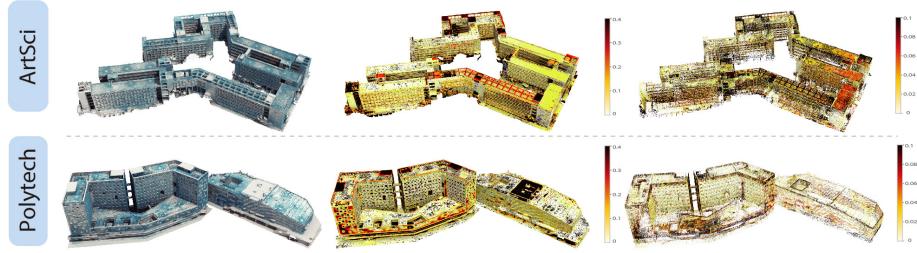


Fig. 4: The visualization of scanned point clouds and reconstruction error maps for real scene ArtSci and Polytech.

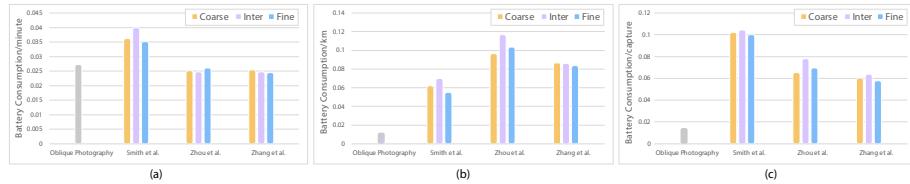


Fig. 5: Battery consumption of different methods with different proxies on the real scene ArtSci. (a): battery consumption per minute (%/minute) ↓; (b): battery consumption per Km ( $10^{-3}\%/\text{km}$ ) ↓; (c): battery consumption between two captures ( $10^{-2}\%$ ) ↓.

For real scene Polytech, the overall error of registration is  $6\text{mm}$ ; for real scene ArtSci, the overall error of registration is  $3\text{mm}$ .

Fig. 4 shows the scanned point clouds, the accuracy maps, and the completeness maps for both the real scene ArtSci and Polytech.

## 5.2 UAV Capturing Cost

Along with the length of the flight path, the efficiency of the path, such as the total turning angles, affects the overall energy consumption. The drone consumes more energy when it accelerates and decelerates near the turns.

Fig. 5 shows the battery consumption statistics of the flight paths planned with different methods on the real scene ArtSci. As we can see in this figure, the capturing cost, or the efficiency of the flight path, is mainly affected by the path pattern. Since oblique photography has the simplest path (Fig. 2), it has nearly the lowest battery cost. Generally, the method proposed by Zhang et al. [44] has a lower battery consumption than the other two methods, since they explicitly optimize the path efficiency in their cost function.

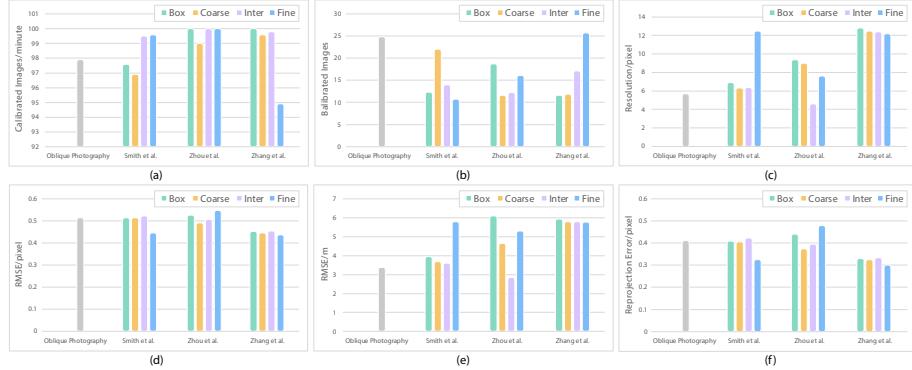


Fig. 6: Aerotriangulation error of different methods with different proxies on the synthetic scenes Town. (a): calibrated images per minute (#/m) ↑; (b): the rate of successfully calibrated images (%) ↑; (c): average resolution per pixel ( $1e^{-3}m/pixel$ ) ↓; (d): root mean square error in pixel ( $1e^{-3}pixel$ ) ↓; (e): root mean square error in meter ( $1e^{-3}m$ ) ↓; (f): reprojection error (pixel) ↓.

### 5.3 Aerotriangulation Error

Before reconstructing the scenes, a process named aerotriangulation, which is a triangulation with aerial images is first performed on the captured images to determine the pose of the cameras and to obtain a sparse point cloud of the environment.

Fig. 6 shows the statistics of the aerotriangulation error on different proxies with the overlap as 90% tested on the scene Town. As indicated in Fig. 6, the aerotriangulation results of the method proposed by Smith et al. [37] and the method proposed by Zhou et al. [46] are quite sensitive to the different levels of proxies. For the method proposed by Zhang et al. [44], the RMS-pixel (root mean square error in pixel), RMS-meter (root mean square error in meter), and the reprojection error decrease as the proxy go finer. However, compared to Zhou et al. [46], some images captured with this method are not calibrated well. The aerotriangulation results of Zhou et al. [46] and Zhang et al. [44] have lower RMSE in meter than Smith et al. [37]. The images captured by oblique photography are also well-calibrated and have a low aerotriangulation error as the images are overlapped with each other strictly.

### 5.4 Reconstruction Accuracy and Completeness

We evaluate the reconstruction results of the four planners with different proxies and the results of reconstruction accuracy and completeness are shown in Fig. 7. The evaluation is performed on scene School with 90% overlap.

The value of 90% or 95% accuracy  $x$  means that for all the closest points of the vertices of the reconstruction model on the ground-truth model, 90% or

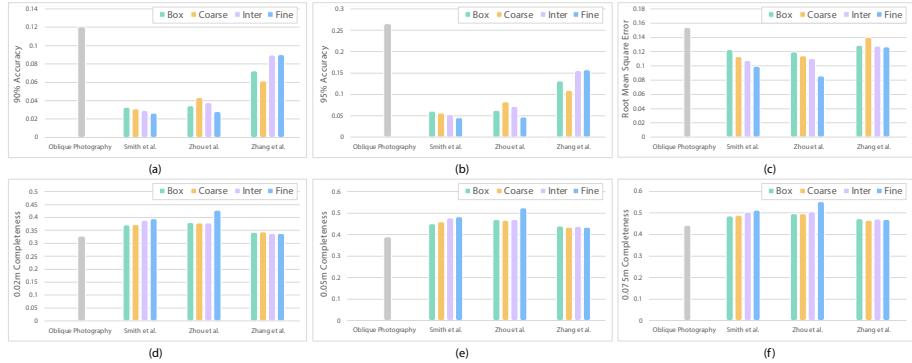


Fig. 7: Reconstruction error of different methods with different proxies on the synthetic scene School. (a): 90% accuracy ( $m \downarrow$ ); (b): 95% Accuracy ( $m \downarrow$ ); (c): root mean square error ( $m \downarrow$ ); (d): 0.02m completeness (%)  $\uparrow$ ; (e): 0.05m completeness (%)  $\uparrow$ ; (f): 0.075m completeness (%)  $\uparrow$ .

95% of them have a distance less than  $x$ . The value of  $0.02m$ ,  $0.075m$ , or  $0.075m$  completeness  $x\%$  means that for all the closest points of the vertices of the ground-truth model on the reconstruction model,  $x\%$  of them have a distance less than  $0.02m$ ,  $0.05m$ , or  $0.075m$ . A smaller 90% and 95% accuracy value mean higher accuracy and a larger  $0.02m$ ,  $0.05m$ , or  $0.075m$  completeness value means higher completeness.

As shown in Fig. 7, both the accuracy and the completeness of reconstructed results of the methods proposed by Smith et al. [37] and Zhou et al. [46] increase as the proxy goes finer. The accuracy and completeness of the results by Zhang et al. [44] are not quite consistent with the proxy.

The visualization of the reconstructed results and reconstruction error is shown in Fig. 8. The values for accuracy and completeness are clamped to  $0 \sim 0.04$  and  $0 \sim 0.1$ . As one could expect, the complex geometry and high occlusion induce lower accuracy and completeness of reconstruction.

In general, the method proposed by Smith et al. [37] and Zhou et al. [46] get higher accuracy and completeness compared to oblique photography and Zhang et al. [44]. However, the paths produced by the method proposed by Zhang et al. [44] have higher quality and consume less energy.

## 5.5 Comparison of Different Planners

The path generated with oblique photography simply follows a Zig-Zag pattern thus the target scene is completely covered and the captured images are well calculated. As the baseline planner, oblique photography has the lowest energy cost among all the four planners but results in a roughest reconstruction. The reconstruction error mainly comes from the occlusion between different buildings and other objects since it can not dive into the space between them. Both the

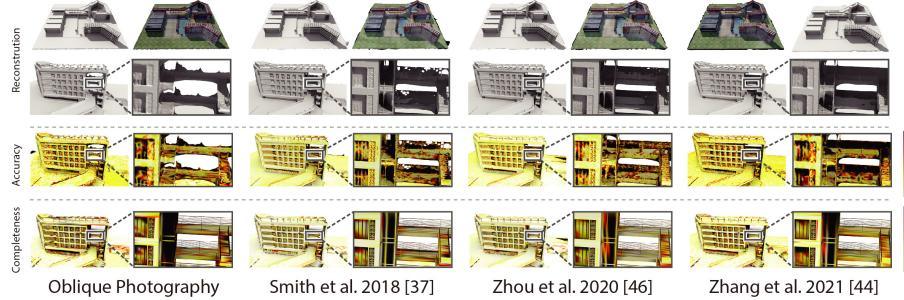


Fig. 8: Visual comparisons of the resulting 3D reconstruction and the corresponding reconstruction error produced by different methods. A higher value means lower accuracy and less completeness for the second and the third rows.

method proposed by Smith et al. [37] and Zhou et al. [46] get a much higher quality reconstruction than oblique photography. In general, the method proposed by Zhou et al. [46] cost less energy compared to the Smith et al. [37]. For the method proposed by Zhang et al. [44], although they get a higher reconstruction error than Smith et al. [37] and Zhou et al. [46], the battery consumption is reduced due to the continuity of the generated paths.

## 6 Simulator and Applications

Although there are 3D instance segmentation datasets, e.g., S3DIS [1], ScanNet [11] and SceneNN [18], they are all collected from indoor scenes and still not enough for deep learning based methods. Please note that there is basically no decent dataset for learning 3D building instance segmentation in spacious outdoor scenes, especially for complicated urban regions.

In this context, our released UrbanScene3D provides rich, large-scale urban scene building annotation data for 3D instance segmentation research. To segment and label 3D architectures, we manually extract all single building models from the entire scene model. Every building is then assigned a unique label, forming an instance segmentation map; see the top right of Fig. 9 for an example. The 3D textured models with instance segmentation labels in UrbanScene3D allow users to obtain all kinds of data they would like to have: instance segmentation map, depth map in arbitrary resolution, 3D point cloud/mesh in both visible and invisible places, etc. UrbanScene3D also offers 4K captured aerial videos in some specific real scenes aimed for 3D reconstruction; see the top left of Fig. 9. Together with high-precision laser scans as ground-truth, these data can be effectively used to train and evaluate various SLAM algorithms.

Moreover, with UrbanScene3D, users can also simulate the robots (cars or drones) to test a variety of autonomous tasks in the proposed city environments. The gravity, inertia and collision can be handled by the physical engine of Airsim. Thus, users can easily generate highly realistic data for many tasks, such as depth



Fig. 9: UrbanScene3D also provides the building instance ID for each environment (top left), 4K aerial videos that are aimed at the real scene acquisition (top right), and a simulator built on Unreal Engine and AirSim (bottom).

estimation, autonomous navigation, and novel view synthesis. Meanwhile, both the lighting condition and the weather of each urban scene can be manipulated by users too. That is, users are able to simulate a rainy night campus or a foggy morning campus; see e.g., the bottom of Fig. 9. Such data endowed with large diversity would reduce the discrepancy between the simulated and real-world environments, and hence increase the generalization of proposed algorithms.

## 7 Conclusion and Future Work

We present a large-scale dataset, *UrbanScene3D*, which offers rich data annotations and a wide variety of observations of six representative environments. The corresponding reconstruction results and the ground-truth models/scans can be used to evaluate path planning and MVS algorithms. Besides, the proposed simulator allows users to further explore and capture urban scenes in various data patterns with different lighting/weather conditions. The release of UrbanScene3D would largely benefit the community.

In the future, we plan to do high-level geometric descriptions in the dataset, such as 3D structural points, cross-sectional profiles, wire-frames, or plane segments, etc., to support further research in both computer vision and computer graphics. UrbanScene3D will constantly grow to make more contributions to the data-driven study.

**Acknowledgements.** This work was supported in parts by NSFC (62161146005, U21B2023, U2001206), GD Talent Program (2019JC05X328), DEGP Key Project (2018KZDXM058, 2020SFKC059), Shenzhen Science and Technology Program (RCJC20200714114435012, JCYJ20210324120213036), and Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ).

## References

1. Armeni, I., Sax, A., Zamir, A.R., Savarese, S.: Joint 2D-3D-semantic data for indoor scene understanding. arXiv preprint arXiv:1702.01105 (2017)
2. Behley, J., Garbade, M., Milioto, A., Quenzel, J., Behnke, S., Stachniss, C., Gall, J.: SemanticKITTI: A dataset for semantic scene understanding of lidar sequences. In: Proc. Int. Conf. on Computer Vision. pp. 9297–9307 (2019)
3. Bergmann, P., Meinhardt, T., Leal-Taixe, L.: Tracking without bells and whistles. In: Proc. IEEE Int. Conf. on Pattern Recognition. pp. 941–951 (2019)
4. Brunel, A., Bourki, A., Strauss, O., Demonceaux, C.: FLYBO: A unified benchmark environment for autonomous flying robots. In: Int. Conf. on 3D Vision. pp. 1420–1431 (2021)
5. Chaquet, J.M., Carmona, E.J., Fernández-Caballero, A.: A survey of video datasets for human action and activity recognition. Computer Vision and Image Understanding pp. 633–659 (2013)
6. Chen, K., de Vicente, J.P., Sepulveda, G., Xia, F., Soto, A., Vázquez, M., Savarese, S.: A behavioral approach to visual navigation with graph localization networks. In: Proc. Robotics: Science and Systems. pp. 1–10 (2019)
7. Chen, Z., Tagliasacchi, A., Zhang, H.: BSP-Net: generating compact meshes via binary space partitioning. Proc. IEEE Conf. on Computer Vision & Pattern Recognition pp. 45–54 (2020)
8. Chen, Z., Zhang, H.: Learning implicit fields for generative shape modeling. Proc. IEEE Conf. on Computer Vision & Pattern Recognition pp. 5939–5948 (2019)
9. Choy, C.B., Xu, D., Gwak, J., Chen, K., Savarese, S.: 3D-R2N2: a unified approach for single and multi-view 3D object reconstruction. ECCV pp. 628–644 (2016)
10. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proc. IEEE Conf. on Computer Vision & Pattern Recognition. pp. 3213–3223 (2016)
11. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In: Proc. IEEE Conf. on Computer Vision & Pattern Recognition. pp. 2432–2443 (2017)
12. Du, D., Qi, Y., Yu, H., Yang, Y., Duan, K., Li, G., Zhang, W., Huang, Q., Tian, Q.: The unmanned aerial vehicle benchmark: Object detection and tracking. In: Proc. Euro. Conf. on Computer Vision Workshops. pp. 370–386 (2018)
13. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? The KITTI vision benchmark suite. In: Proc. IEEE Conf. on Computer Vision & Pattern Recognition. pp. 3354–3361 (2012)
14. Groueix, T., Fisher, M., Kim, V.G., Russell, B.C., Aubry, M.: AtlasNet: a Papier-Mâché approach to learning 3D surface generation. Proc. IEEE Conf. on Computer Vision & Pattern Recognition pp. 216–224 (2018)
15. Gupta, S., Davidson, J., Levine, S., Sukthankar, R., Malik, J.: Cognitive mapping and planning for visual navigation. In: Proc. IEEE Conf. on Computer Vision & Pattern Recognition. pp. 2616–2625 (2017)
16. He, K., Gkioxari, G., Dollár, P., Girshick, R.B.: Mask R-CNN. In: Proc. Int. Conf. on Computer Vision. pp. 2980–2988 (2017)
17. Hepp, B., Nießner, M., Hilliges, O.: Plan3D: Viewpoint and trajectory optimization for aerial multi-view stereo reconstruction. ACM Trans. on Graphics pp. 4:1–4:17 (2018)

18. Hua, B.S., Pham, Q.H., Nguyen, D.T., Tran, M.K., Yu, L.F., Yeung, S.K.: SceneNN: a scene meshes dataset with aNNotations. In: Int. Conf. on 3D Vision. pp. 92–101 (2016)
19. Knapitsch, A., Park, J., Zhou, Q.Y., Koltun, V.: Tanks and temples: Benchmarking large-scale scene reconstruction. ACM Trans. on Graphics (Proc. SIGGRAPH) pp. 78:1–78:13 (2017)
20. Koch, T., Körner, M., Fraundorfer, F.: Automatic and semantically-aware 3D UAV flight planning for image-based 3D reconstruction. Remote Sensing p. 1550 (2019)
21. Liu, J., Ji, S.: A novel recurrent encoder-decoder structure for large-scale multi-view stereo reconstruction from an open aerial dataset. In: Proc. IEEE Conf. on Computer Vision & Pattern Recognition. pp. 6050–6059 (2020)
22. Liu, Y., Cui, R., Xie, K., Gong, M., Huang, H.: Aerial path planning for online real-time exploration and offline high-quality reconstruction of large-scale urban scenes. ACM Trans. on Graphics (Proc. SIGGRAPH Asia) pp. 226:1–226:16 (2021)
23. Liu, Y., Xie, K., Huang, H.: VGF-Net: Visual-geometric fusion learning for simultaneous drone navigation and height mapping. Graphical Models pp. 101108:1–101108:9 (2021)
24. Luo, X., Huang, J., Szeliski, R., Matzen, K., Kopf, J.: Consistent video depth estimation. ACM Trans. on Graphics (Proc. SIGGRAPH) pp. 71:1–71:13 (2020)
25. Lyu, Y., Vosselman, G., Xia, G.S., Yilmaz, A., Yang, M.Y.: UAVid: A semantic segmentation dataset for UAV imagery. ISPRS J. Photogrammetry and Remote Sensing pp. 108–119 (2020)
26. Mandal, M., Kumar, L.K., Vipparthi, S.K.: MOR-UAV: A benchmark dataset and baselines for moving object recognition in UAV videos. In: Proc. ACM Conf. on Multimedia. pp. 2626–2635 (2020)
27. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: ECCV (2020)
28. Mo, K., Zhu, S., Chang, A.X., Yi, L., Tripathi, S., Guibas, L.J., Su, H.: PartNet: A large-scale benchmark for fine-grained and hierarchical part-level 3D object understanding. In: Proc. IEEE Conf. on Computer Vision & Pattern Recognition. pp. 909–918 (2019)
29. Mueller, M., Smith, N., Ghanem, B.: A benchmark and simulator for UAV tracking. In: Proc. Euro. Conf. on Computer Vision. pp. 445–461 (2016)
30. Nathan Silberman, Derek Hoiem, P.K., Fergus, R.: Indoor segmentation and support inference from RGBD images. In: Proc. Euro. Conf. on Computer Vision. pp. 746–760 (2012)
31. Perera, A.G., Wei Law, Y., Chahl, J.: UAV-GESTURE: A dataset for UAV control and gesture recognition. In: Proc. Euro. Conf. on Computer Vision Workshops. pp. 0–0 (2018)
32. Pisharady, P.K., Saerbeck, M.: Recent methods and databases in vision-based hand gesture recognition: A review. Computer Vision and Image Understanding pp. 152–165 (2015)
33. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: PointNet++: deep hierarchical feature learning on point sets in a metric space. Proc. Conf. on Neural Information Processing Systems pp. 5099–5108 (2017)
34. Roberts, M., Dey, D., Truong, A., Sinha, S., Shah, S., Kapoor, A., Hanrahan, P., Joshi, N.: Submodular trajectory optimization for aerial 3D scanning. In: Proc. Int. Conf. on Computer Vision. pp. 5324–5333 (2017)

35. Ros, G., Sellart, L., Materzynska, J., Vazquez, D., Lopez, A.M.: The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In: Proc. IEEE Conf. on Computer Vision & Pattern Recognition. pp. 3234–3243 (2016)
36. Schmid, K., Hirschmüller, H., Dömel, A., Grix, I., Suppa, M., Hirzinger, G.: View planning for multi-view stereo 3D reconstruction using an autonomous multicopter. *J. Intelligent & Robotic Systems* pp. 309–323 (2012)
37. Smith, N., Moehrle, N., Goesele, M., Heidrich, W.: Aerial path planning for urban scene reconstruction: a continuous optimization method and benchmark. *ACM Trans. on Graphics (Proc. SIGGRAPH Asia)* pp. 183:1–183:15 (2018)
38. Song, X., Wang, P., Zhou, D., Zhu, R., Guan, C., Dai, Y., Su, H., Li, H., Yang, R.: ApolloCar3D: A large 3D car instance understanding benchmark for autonomous driving. In: Proc. IEEE Conf. on Computer Vision & Pattern Recognition. pp. 5447–5457 (2019)
39. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., et al.: Scalability in perception for autonomous driving: Waymo open dataset. In: Proc. IEEE Conf. on Computer Vision & Pattern Recognition. pp. 2446–2454 (2020)
40. Xia, F., R. Zamir, A., He, Z.Y., Sax, A., Malik, J., Savarese, S.: Gibson Env: real-world perception for embodied agents. In: Proc. IEEE Conf. on Computer Vision & Pattern Recognition. pp. 9068–9079 (2018)
41. Xu, Q., Wang, W., Ceylan, D., Mech, R., Neumann, U.: DISN: deep implicit surface network for high-quality single-view 3D reconstruction. *Proc. Conf. on Neural Information Processing Systems* pp. 490–500 (2019)
42. Yao, Y., Luo, Z., Li, S., Zhang, J., Ren, Y., Zhou, L., Fang, T., Quan, L.: Blend-edMVS: A large-scale dataset for generalized multi-view stereo networks. In: Proc. IEEE Conf. on Computer Vision & Pattern Recognition. pp. 1790–1799 (2020)
43. Yin, W., Zhang, J., Wang, O., Niklaus, S., Mai, L., Chen, S., Shen, C.: Learning to recover 3D scene shape from a single image. In: Proc. IEEE Conf. on Computer Vision & Pattern Recognition. pp. 204–213 (2021)
44. Zhang, H., Yao, Y., Xie, K., Fu, C.W., Zhang, H., Huang, H.: Continuous aerial path planning for 3D urban scene reconstruction. *ACM Trans. on Graphics (Proc. SIGGRAPH Asia)* pp. 225:1–225:15 (2021)
45. Zhang, X., Zhao, P., Hu, Q., Ai, M., Hu, D., Li, J.: A UAV-based panoramic oblique photogrammetry (POP) approach using spherical projection. *J. Photogrammetry and Remote Sensing* pp. 198–219 (2020)
46. Zhou, X., Xie, K., Huang, K., Liu, Y., Zhou, Y., Gong, M., Huang, H.: Offsite aerial path planning for efficient urban scene reconstruction. *ACM Trans. on Graphics (Proc. SIGGRAPH Asia)* pp. 192:1–192:16 (2020)
47. Zhou, X., Yi, Z., Liu, Y., Huang, K., Huang, H.: Survey on path and view planning for UAVs. *Virtual Reality & Intelligent Hardware* pp. 56–69 (2020)
48. Zhou, Y., Huang, J., Dai, X., Luo, L., Chen, Z., Ma, Y.: HoliCity: a city-scale data platform for learning holistic 3D structures. *arXiv preprint arXiv:2008.03286* (2020)
49. Zhu, P., Du, D., Wen, L., Bian, X., Ling, H., Hu, Q., Peng, T., Zheng, J., Wang, X., Zhang, Y., et al.: Visdrone-vid2019: The vision meets drone object detection in video challenge results. In: Proc. Int. Conf. on Computer Vision Workshops. pp. 1–9 (2019)
50. Zhu, Y., Mottaghi, R., Kolve, E., Lim, J.J., Gupta, A., Fei-Fei, L., Farhadi, A.: Target-driven visual navigation in indoor scenes using deep reinforcement learning. In: Proc. IEEE Int. Conf. on Robotics & Automation. pp. 3357–3364 (2017)