A Study of Aerial Image-Based 3D Reconstructions in a Metropolitan Area

Susana Ruano(✉) and Aljosa Smolic

V -SENSE, Trinity College Dublin, Dublin, Ireland
{ruanosas,smolica}@tcd.ie

Abstract. We present a study of image-based 3D reconstruction methods that use aerial images in large-scale metropolitan areas. Specifically, the study analyzes both open-source methods from the state of the art, and some of the most used commercial photogrammetry applications. The performance of these methods is measured against the densest annotated LiDAR point cloud available at a city scale. The study not only analyzes the accuracy and completeness of the reconstruction methods in the metropolitan area as in previous studies, but it also evaluates their performance for different city elements such as trees, windows, roofs, etc. which are present in the annotated ground truth model.

Keywords: Study · 3D Reconstruction · Structure-from-motion · Multi-view stereo

1 Introduction

Urban environments have been extensively studied for many purposes such as path planning, to improve the acquisition of images, to create 3D reconstructions [27], or to perform semantic segmentation of roads in the city using RGB-thermal images [29]. The creation of datasets and benchmarks focused on metropolitan areas is therefore an interesting topic and new classified point clouds of cities and datasets of multimodal aerial sources are being released, both for semantic segmentation [3], and also to perform path planning [21]. These recent publications show that despite the progress in computer vision techniques there is still a need for a common framework to evaluate them under the same conditions in different domains.

It is common knowledge that especially image-based 3D reconstruction techniques have been struggling to find the appropriate ground-truth data to evaluate the algorithms. Numerous research efforts were done to provide a successful solution to this by creating benchmarks: since the earliest works [19] which provide a few samples to evaluate the algorithms, to the more recent and complete benchmarks which cover multiple scenarios [7,18]. However, it is still challenging to have a great extension of the city covered with enough accuracy. Generally, specific equipment such as a LiDAR is needed to collect the ground truth making

A. A. de Sousa et al. (Eds.): VISIGRAPP 2022, CCIS 1815, pp. 304–323, 2023.
https://doi.org/10.1007/978-3-031-45725-8_15

it unaffordable for most people. In particular, to cover a large extension of an urban environment, not only the equipment but also the planning of a mission (e.g., flight of an helicopter) is a barrier for creating new datasets.

A metropolitan area can accommodate a great variety of elements: buildings, roads, parks, etc. Its diversity can present a challenge for 3D reconstruction techniques and the ground-truth data available does not typically provide detailed information about these elements the city contains. However, in the same way that general purpose 3D reconstruction benchmarks include a great variety of scenarios, a benchmark for the study of an urban environment would benefit from having this type of information. Previous studies [30] remark the importance of including the categorization of the city to support the observations made about the quality of the reconstructions.

The creation of 3D models from a collection of images has been studied for many years [2,4,10,16,17,22,25] and it is a fundamental problem of computer vision [5]. The techniques to create 3D models can be considered specific cases, such as aerial images of a mountainous area [14] which can be later used in augmented reality applications [13], but they are normally general purpose methods. A complete pipeline to obtain an image-based reconstruction includes two different stages: Structure-from-Motion (SfM) and Multi-View Stereo (MVS). The former, to recover a sparse model and the camera poses, and the later, to obtain a detailed dense point cloud. There are very popular open-source techniques which are normally used in the 3D reconstruction studies [23], but there are also some commercial solutions with different licenses that can be used to obtain 3D models from images, which are not included in the studies very often.

In this paper, we present an extension of the previous work [15] with an evaluation for image-based 3D reconstruction pipelines not limited to combination of open-source techniques. We expand the previous work by including commercial applications in the evaluation and thus improving the study of a metropolitan area. We use the annotated point cloud of DublinCity [30] and make use of the initial study done in [15] where the city is analyzed. The analysis not only includes the evaluation at scene level but also per category of urban element, as it was done in the benchmark.

2 Previous Work

Benchmarks have been used to evaluate image-based reconstructions for decades. The first attempts were limited in various aspects [19]: they have a limited number of ground-truth models (only two), they were focused only in a part of the 3D reconstruction pipeline (MVS), and the ground-truth was acquired in a very controlled environment. With time, these type of benchmarks have been enhanced with different sets of images taken under specific lighting conditions and with additional 3D ground-truth models [1]. Nevertheless, the limitation of having all the images acquired in a controlled setting was not tackled until the EPFL benchmark was released [24]. In that work, the ground truth provided was acquired with a terrestrial LiDAR. Since then, outdoor scenarios started

to be considered in the evaluations. However, as it is the pioneer for outdoor environments, it has the same drawback as the very first indoor benchmarks: the variety of models captured for the benchmark was small and the coverage of the scene was narrow. Still, it was a significant progress.

The limitation regarding the variety of models and scenarios was overcome with two other benchmarks released at a very similar time: Tanks and Temples [7] and the ETH3D benchmark [18]. The former is used to evaluate the output of the complete image-based 3D reconstruction pipeline (i.e., the dense point cloud) and instead of providing a collection of images they give as input a video. The later is focused on the MVS stage, and therefore, it provides the camera calibration and poses along with the set of images. A terrestial LiDAR was used in these benchmarks to capture the ground-truth, and in some scenes including buildings, some details such as the roofs were not covered.

Other types of LiDAR that can be used to acquire the ground truth are: 3D Mobile Laser Scanning (MLS) and Aerial Laser Scan (ALS). This type of equipment is the one typically used to acquire ground-truth data that covers large metropolitan areas. Among city benchmarks, some of them do not include images alongside such as the the TerraMobilita/iQmulus benchmark [26], Paris-rue-madame dataset [20] and the Oakland 3D Point Cloud dataset [11]. Others, do not use the images to create the 3D reconstructions (like in the Toronto/Vaihingen ISPRS benchmark used in [28]), instead the point cloud is used. Some of the benchmarks that include images such as the Kitti benchmark [8], and the ISPRS Test project on Urban Classification and 3D Building Reconstruction [12] are not focused on the evaluation of the MVS output; instead, they focus on the evaluation of other tasks such as stereo, object detection, or reconstructions made from roofs.

Some of the benchmarks cited above have an annotated point cloud of the city [11,20,26], but as pointed out, they do not provide images, with the exception of the ISPRS Test project on Urban Classification and 3D Building Reconstruction [12] and the Ai3Dr Benchmark [15]. The main difference between these last two regarding the ground-truth model is the density. The former is about six points per m^2 and the later around 350 points per m^2. Also, the later uses the a very rich annotated point cloud [30], and two different types of images, as opposed to a single set as in the former. Our work is an extension of the evaluation presented in the Ai3Dr Benchmark, including an extension of the pipelines evaluated under the same conditions.

Fig. 1. Areas under evaluation. The areas of the city considered in this study are highlighted in the map with a top view of the ground truth.

3 Urban Environment

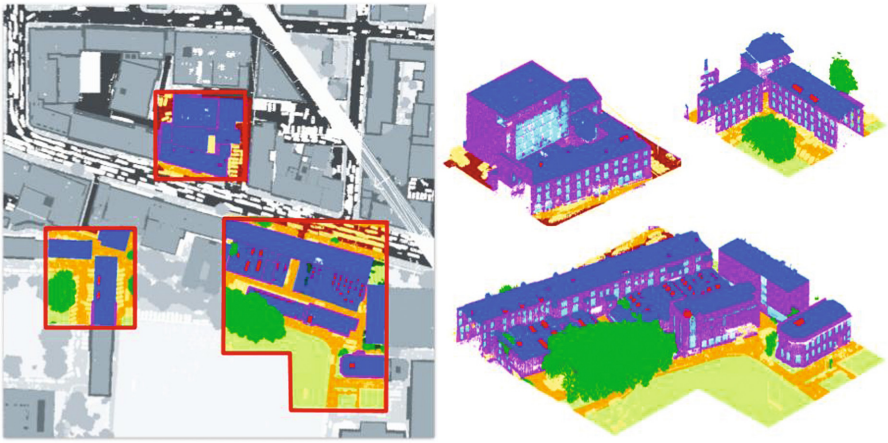


Fig. 2. Hidden areas. On the left, an example of a hidden area (in color) over the whole region (in grey). On the right, the hidden area represented in 3D [15].

Table 1. Number of points and percentage of points per class in each evaluated area. Also the number of points and percentage used as the hidden zone [15].

		undefined	facade	window	door	roof	r. window	r. door	sidewalk	street	grass	tree
Area 1	# points ($\times 10^3$)	1397	3048	540	40	10727	481	3	2029	3210	1587	3471
	percentage	5.27	11.49	2.04	0.15	40.43	1.81	0.01	7.65	12.10	5.98	13.08
	hidden # p. ($\times 10^3$)	1015	1816	312	27	2849	247	2	1602	2441	1353	2720
	hidden pct	7.06	12.63	2.17	0.19	19.81	1.72	0.01	11.14	16.97	9.41	18.91
Area 2	# points ($\times 10^3$)	1615	2776	308	40	7807	112	3	2789	2642	4061	2852
	percentage	6.46	11.10	1.23	0.16	31.22	0.45	0.01	11.15	10.57	16.24	11.41
	hidden # p. ($\times 10^3$)	1262	1547	220	27	2708	47	1	2010	1993	2226	1897
	hidden pct	9.05	11.105	1.58	0.19	19.43	0.34	0.01	14.42	14.30	15.97	13.61

The metropolitan area evaluated in this paper is the city of Dublin. In particular, the same extension used in the Ai3Dr benchmark [15] is considered for the evaluation. As it is explained there, two different regions of the city were selected to be evaluated, which include a variety of buildings, streets and parks with diverse styles, sizes and distribution, from the Trinity College campus to modern buildings as well as different structures such as trail tracks. These regions are depicted in Fig. 1 and we refer to them as *Area 1* and *Area 2* from now on.

3.1 Ground Truth

A selection of the DublinCity [30] annotated point cloud is used as ground truth for this evaluation, which includes a representative part of the city. The areas were carefully analyzed from the whole extension of the city in [15], discarding the ones with the following characteristics: small extension (less than 250×250 m^2 of the city), those areas with a low percentage of points in certain categories such as trees or grass, tiles that contain elements which are temporarily in the city and can degrade the performance (e.g., cranes) and those ones which have less that 90% of points classified.

Fig. 3. Oblique and nadir images. A sample of oblique images (left) and nadir ones (right).

Accordingly, the representative areas *Area 1* and *Area 2* correspond to these selected pieces of ground truth. This ensures a balanced distribution of points in each category while avoiding parts which can diminish the evaluation. It also guarantees the content is representative and diverse enough to make an evaluation per urban object category. Lastly, the magnitude of the city to be analyzed is also adequate so the algorithms process them in a reasonable amount of time.

As explained in [15], we are also using *hidden areas* for the evaluation (see and example in Fig. 2). These consist of meaningful sections of a particular area, selected to preserve the meaningfulness of the class distribution for the evaluation. However, the specific sections of the ground truth are not disclosed to the final user to avoid fine tuning to a specific region during the online evaluation.

3.2 Image Sets

We use three different sets of images to build the 3D models of each area: oblique, nadir, and a combination of both. Images in the oblique and nadir sets come from the groups described in [30] (Fig. 3 illustrates a sample of each of these images). As the original dataset contains a significantly large amount of images, both oblique and nadir, we use the selection done in [15]. Therefore, we use only the ones that will have a meaningful contribution to the 3D reconstruction of each of our areas. For the selection, COLMAP's SfM algorithm [16] was used with the complete image datasets, and every image that does not contribute to at least 1500 3D points was discarded. This is was done for both oblique and nadir image sets.

3.3 Urban Categorization

Table 1 illustrates how the 3D points in the ground-truth models are distributed, with respect to their class, both for the complete models and also the hidden areas. The table shows that roof is the class with higher point count, and also its percentage is more balanced in the hidden areas with respect to other classes. The class door is the one with the lowest point count, and the undefined data, which could potentially introduce errors or other inaccuracies, represents a maximum of 9% of the points of all areas. Other relevant metrics are that *Area 1* has almost twice the number of window points than *Area 2* and four times of roof windows; while *Area 2* has three times more grass points.

4 Experimental Setup

The goal of this study is to evaluate the performance of image-based 3D reconstruction techniques in the metropolitan area described in Sect. 3. The inputs of the 3D reconstruction processes are three different set of aerial images (oblique, nadir and combined) and the GPS data captured in the flight. This information is given in the EPSG 29902 reference system in separate files, and, in particular,

the nadir set has also geographic information embedded as Exif metadata. All this information is available in the Ai3Dr benchmark. As this study is an extension of [15], we analogously evaluate the final result of the 3D reconstruction process (i.e., the final dense point cloud generated by the methods) and not the intermediate stages (SfM or MVS) separately.

The reasons to choose this approach were essentially two. Firstly, make it adaptable to new approaches that may not follow the typical steps of the pipeline. Secondly, as the measurements of the camera positions available are based on GPS and no additional information (e.g., the orientation of the camera) is given, these measurements are used only as a coarse approximation and are not indicated to be used as ground truth for evaluating intermediate steps.

4.1 3D Reconstruction Techniques

We can divide the 3D reconstruction techniques evaluated in two groups: open-source techniques and commercial applications. On the one hand, the former techniques are freely available, they are usually focused on one specific step of the process but some of them can be mixed to complete the pipeline. Users have access to the details in the code, they are typically used to solve general purpose reconstruction problems, and they have multiples parameters to be configured to adapt the algorithms to specific cases. Commercial applications are in general prepared for a specific type of reconstructions, and they have free trials available but later one needs to pay for the software. The details of the algorithms used are not disclosed, there are intermediate steps of refinement, they have usable interfaces, and default parameters are well adjusted for the tasks. Also, they are fast and are well optimized.

We are combining several open-source SfM and MVS algorithms to create the open-source 3D reconstruction pipelines. Firstly, for SfM, we use COLMAP [16], which includes a geometric verification strategy that helps improving robustness on both initialization and triangulation, and includes an improved bundle adjustment algorithm with outlier filtering. Furthermore, we use two different SfM approaches implemented in OpenMVG: a global one [10] based on the fusion of relative motions between image pairs, and an incremental one [9] that iteratively adds new estimations to the initial reconstruction minimizing the drift with successive steps of non-linear refinement. Secondly, for MVS, we also use COLMAP's approach [17] that jointly estimates depth and normal information and makes a pixel-wise view selection using photometric and geometric priors. Moreover, we also use OpenMVS [17] which does efficient patch-based stereo matching followed by a depth-map refinement process.

There are several commercial applications that are used to create image-based 3D reconstructions in professional environments. We evaluate three of the most popular applications: Agisoft Metashape[1], Pix4D[2] and RealityCapture[3].

[1] https://www.agisoft.com/.
[2] https://www.pix4d.com/.
[3] https://www.capturingreality.com/.

Agisoft Metashape is a stand-alone photogrammetry software that generates 3D models from a colletion of images and they can be used for cultural heritage works, for GIS applications, etc. They cover a large range of specific applications in their professional edition. Among their features we can find digital elevation model generation, dense point cloud editing and multispectral imagery processing. They also offer a python scripting binding and their program can be used in Windows, Linux and MacOs. Pix4D offers a photogrammetry suite specialized in mobile and drone mapping. They offer solutions for inspection, agriculture and surveying. We have selected the Pix4Dmapper product to make the reconstructions, because although Pix4Dmatic is expected to perform photogrammetry at large scale, we found that more information (e.g., orientation of the camera) was needed for initiate the standard procedure. Pix4D is not available for Linux. RealityCapture is a photogrammetry software recently acquired by Epic Games. They offer a solution to different task such as 2D-3D mapping, 3D printing, full body scans, assets for games, etc. They claim to have a long trajectory and recognition in the computer vision since they have created CMPMVS [6] and their researchers are also recongnized in the community. Their application is available for Windows only.

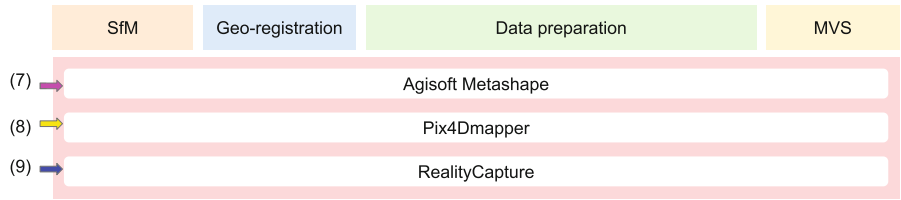


Fig. 4. Scheme of 3D reconstruction pipelines tested. (1) to (6) as in [15], (7) Metashape, (8) Pix4D, (9) RealityCapture.

4.2 Pipelines Under Study

Using the aforemention 3D reconstruction techniques, we list here the pipelines that are tested in this study:

1. COLMAP(SfM) + COLMAP(MVS)
2. COLMAP(SfM) + OpenMVS
3. OpenMVG-g + COLMAP(MVS)
4. OpenMVG-g + OpenMVS
5. OpenMVG-i + COLMAP(MVS)
6. OpenMVG-i + OpenMVS
7. Metashape
8. Pix4Dmapper
9. RealityCapture

As it is depicted in Fig. 4, pipelines (7) to (9), which correspond to commercial applications, are treated as a complete solution from the input to the output. For pipelines (1)-(6), we use the same configuration as in [15], which correspond to the pipelines assembled with open-source techniques and four stages are needed: SfM, geo-registration, data preparation and MVS. Feature detection and matching are done in the SfM step. Then, the GPS information is used to coarsely register the sparse cloud to the ground truth. The following step is a preparation for performing the densification, and the final one is the densification itself. A different approach is followed in the pipelines (7)-(9) because they consist of closed solutions for all the stages of the pipeline. The coarse registration is done establishing EPSG:29902 as coordinate system for the output. The nadir images have enough information to be geo-referenced without using additional data but for the oblique set, the geographic information provided in the benchmark is needed. Depending on the method, the registration can be done in different stages. For example, (7) can calculate the reconstruction in a local coordinate system and then change it when exporting the point cloud, however, (8) requires to have the coordinate estimation beforehand to have good results.

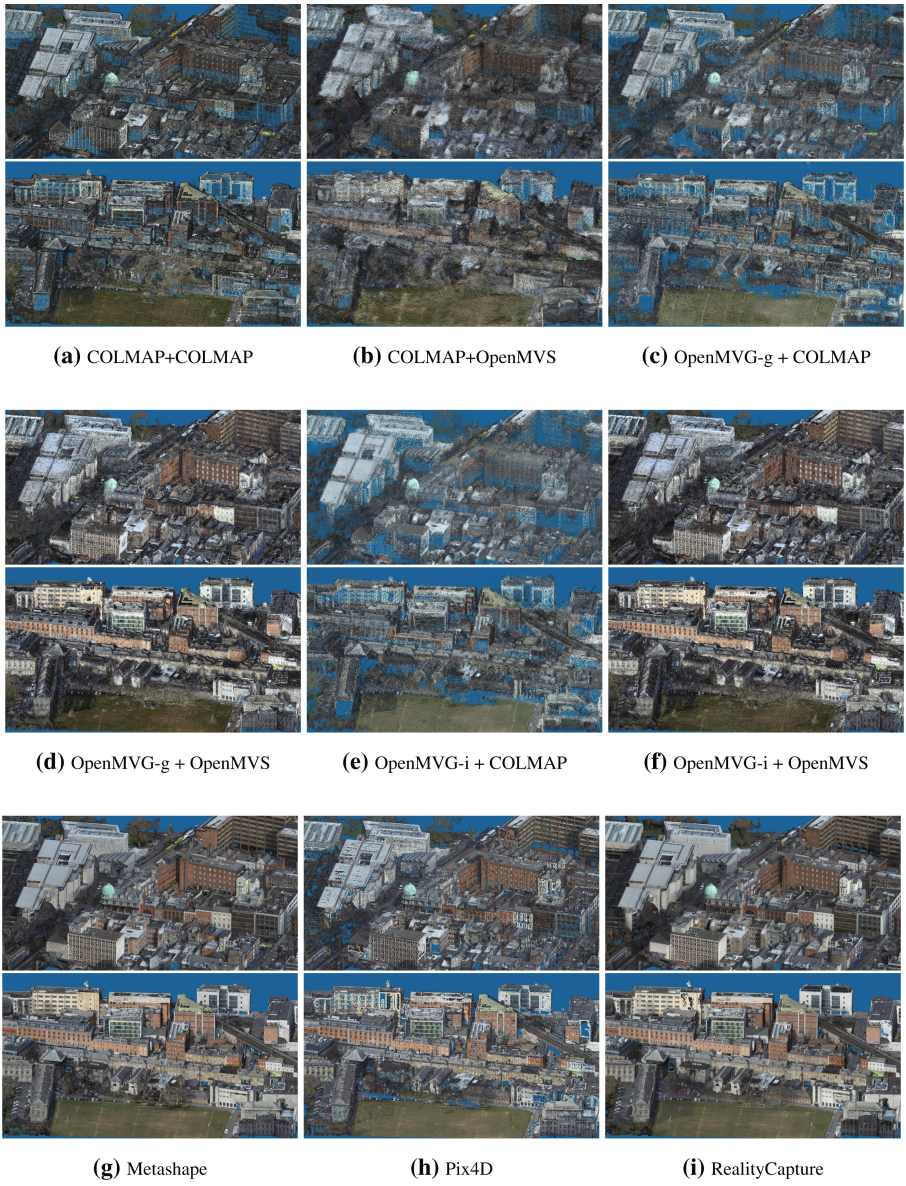
The versions of each software used in this study are :

- COLMAP v3.6
- OpenMVG v1.5
- OpenMVS v1.1
- Agisoft Metashape Pro v1.7.5 for Linux
- Pix4D v4.7.5 for Windows
- Reality Capture v1.2 for Windows

The parameter configurations in the open-source pipelines are dependent on the stage of the pipeline and the method, as indicated in [15]. In general terms, COLMAP's parameters are the same as in DublinCity [30] in all the stages of the pipeline. OpenMVG also uses the default parameters and OpenMVS uses the parameters reported in the ETH3D benchmark. We used the default parameters as well in the Metashape, Reality Capture, and Pix4D.

4.3 Alignment

The output of the aforementioned image-based 3D reconstruction pipelines is coarsely registered to the ground truth thanks to the geographical information. This is the same situation as in [15]. There, it is shown that the coarsely registration needs a refinement process in order to be perfectly aligned. The registration refinement process typically consists of a 7DoF ICP algorithm, a strategy as followed in [7], for example. Schops et al. [18] use a more sophisticated approach using the color information of the laser scan, something that is not available in our case. For our approach, we follow the same strategy as in [15], further refined with a 7DoF ICP process with the point cloud and the ground truth.



(a) COLMAP+COLMAP **(b)** COLMAP+OpenMVS **(c)** OpenMVG-g + COLMAP

(d) OpenMVG-g + OpenMVS **(e)** OpenMVG-i + COLMAP **(f)** OpenMVG-i + OpenMVS

(g) Metashape **(h)** Pix4D **(i)** RealityCapture

Fig. 5. Qualitative 3D reconstruction results. Point clouds obtained with the oblique and nadir images combined in *Area 1* (top) and *Area 2* (bottom).

5 Experimental Results

Following the work in [7,15,30], we use the following measurements to evaluate performance:

– **Precision**, P: measures the accuracy of the reconstruction.

$$P(d) = \frac{|dist_{I \to G}(d)|}{|I|} 100 \tag{1}$$

– **Recall**, R: measures the completeness of the reconstruction.

$$R(d) = \frac{|dist_{G \to I}(d)|}{|G|} 100 \tag{2}$$

– **F score**, F: a combination of both P and R.

$$F(d) = \frac{2P(d)R(d)}{P(d) + R(d)} \tag{3}$$

Where d is a given threshold distance, I is the point cloud under evaluation and G is the ground-truth point cloud. $|\cdot|$ is the cardinality and $dist_{I \to G}(d)$ are the points in I with a distance to G less than d and $dist_{G \to I}(d)$ is analogous (i.e., $dist_{A \to B}(d) = \{a \in A \mid \min_{b \in B} \|a - b\|_2 < d\}$, A and B being point clouds). To perform the evaluation per class, the point under evaluation is assigned the same class as its nearest neighbor in the ground-truth. Although there are plenty of metrics that can be used to measure the quality of the reconstruction algorithms, such as the mean distance to the ground truth, as used in [23], P, R, and therefore F, are more robust to outliers.

For each pipeline, set of images and area (including the hidden parts), we calculate P, R, and F. We use a value of d in the range of 1 cm to 100 cm, which produces better results in every method when we increase the value of d, as an expected result after increasing the tolerance. The results reported in this evaluation use a value of 25 cm for d, similar to [15,30], which represents a good compromise between the limitations of the image resolution and the meaningfulness of the precision, since selecting a very small distance would mean poorer performances for all the methods and with a larger distance the precision would be less informative.

5.1 Scene Level Evaluation

Table 2 shows the evaluation at scene level. This means that all the points in the ground truth are treated in the same way, ignoring to which class they belong. We can see in the results that the reconstructions done with the oblique set achieve the lowest recall values, in comparison with the reconstructions obtained with the other sets of images, in both areas (also in the hidden parts). Therefore, for this set of images, having a good precision value is determinant to achieve a good F score. Among the pipelines created with open-source methods, COLMAP + COLMAP has the best performance for this type of images in both areas. However, some commercial solutions outperform it. Specifically, the best score in *Area 2* is obtained with Pix4D, whereas in *Area 1* it is with Metashape.

In the nadir set, the recall is usually higher than the precision so the accuracy is not as determinant as in the oblique sets to obtain a good F score. These results

suggest that having different camera angles and less coverage of the same parts of the scene (as it is the case in the oblique set but not in the nadir one) makes the recall value decrease while the precision remains similar. As it can be observed, COLMAP + COLMAP is the best pipeline in *Area 1* whereas OpenMVG-i + OpenMVS is the best in *Area 2* (even in the hidden parts), among the open-source methods. Moreover, among the commercial solutions, we also see different winning methods in *Area 1* and *Area2*: Reality Capture and Pix4D, respectively. For the reconstructions obtained with the combined imagery, OpenMVG-g + OpenMVS and OpenMVG-i + OpenMVS are the open-source pipelines with the highest F score: the former in *Area 1* and latter in *Area 2*. This is slightly different for the commercial solutions as they present their best results with the same method for both the nadir and the combined set of images. In fact, the F score of the reconstructions obtained with the nadir and the combined sets of images is much more similar in the commercial applications than in the others. This suggest that the nadir images are very well treated by the commercial solutions. We can also observe that some open-source pipelines have higher recall values in the nadir and combined sets, although this is not enough to beat the commercial ones. However, for the combined set of images, where the open-source techniques have their best results, the difference between this and the worst performing method of the commercial applications is not very significant. 79.36 and 79.5 for OpenMVG-g + OpenMVS and Pix4D, respectively, in *Area 1* and, analogously, 81.13 and 81.53 for OpenMVG-i + OpenMVS and Metashape in *Area 2*.

The qualitative results obtained from the reconstructions with the oblique and nadir images together in Area 1 and Area 2, are shown in Fig. 5. The render was done using the same configuration (e.g., size of the points) to make it comparable. We can observe from these results that the point cloud obtained with COLMAP + COLMAP (Fig. 5 (a)) is sharper than the one obtained with COLMAP + OpenMVS (Fig. 5 (b)) in both areas, in accordance with the precision values (74.89 and 27.93 in Area 1; 76.54 and 24.7 in Area 2). We can also see that the highest precision value in Area 1, 87.87 is obtained with Reality-Capture (Fig. 5 (i)) with a very sharp reconstruction. Moreover, there are also differences in the completeness of the reconstructions: OpenMVG-i + OpenMVS (Fig. 5 (f)) and OpenMVG-g + OpenMVS (Fig. 5 (d)) are denser than the rest, this time among all the methods including the commercial ones in both areas. However, their F score values (79.36 and 76.76 in Area 1; 80.64 and 81.13 in Area 2) confirm that, as it can be appreciated in the images, the commercial software seems to give sharper reconstructions (Fig. 5 (g)-(i)), and therefore the best scores in the open-source pipelines are finally worse than all of their F scores (79.5, 80.68, 80.99 in Area 1 and 81.54, 82.73, 85.06 in Area 2), thanks to the higher values in precision.

5.2 Urban Category Centric Evaluation

Additionally, we present a summary of the same measurements calculated above but this time per urban element category. This summary shows three tables, one

Table 2. Study of urban areas (quantitative). Each row shows the results of a specific 3D reconstruction pipeline giving the precision / recall / F score for d=25cm obtained for the reconstruction in each set of images in each area. The best score for each area and image set is in bold letters and the pipelines are as follows: (1) COLMAP + COLMAP, (2) COLMAP + OpenMVS, (3) OpenMVG-g + COLMAP, (4) OpenMVG-g + OpenMVS, (5) OpenMVG-i + COLMAP, (6) OpenMVG-i + OpenMVS, (7) Metashape, (8) Pix4D, (9) RealityCapture.

	Area 1		
	oblique	nadir	oblique and nadir
(1)	**79.18** / 60.5 / **68.59**	**73.08** / 68.98 / **70.97**	74.89 / 74.15 / 74.52
(2)	22.74 / 28.28 / 25.21	23.96 / 46.23 / 31.57	27.93 / 60.84 / 38.29
(3)	49.42 / 13.09 / 20.69	44.02 / 47.95 / 45.9	48.74 / 58.24 / 53.07
(4)	61.07 / 57.1 / 59.02	56.61 / **73.46** / 63.94	**78.27** / **80.49** / **79.36**
(5)	37.13 / 16.59 / 22.94	36.59 / 43.62 / 39.8	39.48 / 51.36 / 44.64
(6)	55.12 / **64.37** / 59.39	49.14 / 70.75 / 58.0	74.19 / 79.5 / 76.76
(7)	81.6 / **72.5** / **76.78**	84.26 / **76.61** / 80.25	82.91 / **78.56** / 80.68
(8)	77.41 / 49.64 / 60.49	87.23 / 71.41 / 78.53	86.79 / 73.34 / 79.5
(9)	**82.54** / 66.95 / 73.93	**87.67** / 75.82 / **81.32**	**86.87** / 75.86 / **80.99**
	Area 2		
	oblique	nadir	oblique and nadir
(1)	**80.48** / 65.51 / **72.23**	74.97 / 72.34 / 73.63	76.54 / 77.98 / 77.25
(2)	26.85 / 41.81 / 32.7	24.45 / 49.96 / 32.83	24.7 / 63.53 / 35.57
(3)	36.92 / 15.7 / 22.03	33.26 / 36.4 / 34.76	41.94 / 50.3 / 45.74
(4)	40.23 / 54.27 / 46.21	**75.2** / 75.92 / **75.56**	79.3 / 82.03 / 80.64
(5)	38.11 / 15.11 / 21.64	43.76 / 52.0 / 47.52	36.62 / 48.53 / 41.74
(6)	58.52 / **70.43** / 63.92	71.44 / **79.0** / 75.03	**79.77** / **82.54** / **81.13**
(7)	80.96 / **75.24** / 78.0	87.63 / **78.49** / 82.81	83.36 / **79.79** / 81.53
(8)	**94.35** / 67.53 / **78.72**	**94.39** / 76.21 / **84.33**	**93.68** / 77.9 / **85.06**
(9)	83.77 / 73.06 / 78.05	90.37 / 78.32 / 83.92	87.49 / 78.46 / 82.73
	hidden *Area 1*		
	oblique	nadir	oblique and nadir
(1)	**78.68** / 49.89 / **61.06**	**72.69** / 62.77 / **67.37**	74.54 / 68.34 / 71.3
(2)	23.55 / 18.97 / 21.01	24.42 / 36.88 / 29.39	28.38 / 50.45 / 36.32
(3)	43.12 / 6.61 / 11.46	42.52 / 38.01 / 40.14	48.48 / 48.74 / 48.61
(4)	56.57 / 48.7 / 52.34	56.03 / **69.48** / 62.03	**75.36** / **75.76** / **75.56**
(5)	30.83 / 8.88 / 13.79	36.51 / 35.47 / 35.98	38.79 / 42.08 / 40.37
(6)	52.93 / **59.33** / 55.95	48.47 / 67.81 / 56.53	70.56 / 75.02 / 72.72
(7)	79.11 / 65.14 / **71.45**	82.46 / **70.3** / 75.9	80.76 / **73.08** / **76.73**
(8)	78.87 / 36.21 / 49.64	86.19 / 63.61 / 73.2	84.63 / 65.27 / 73.7
(9)	**80.55** / **58.45** / 67.74	**86.25** / 69.37 / **76.89**	**85.39** / 69.53 / 76.65
	hidden *Area 2*		
	oblique	nadir	oblique and nadir
(1)	**80.06** / 61.48 / **69.55**	73.63 / 68.9 / 71.18	75.5 / 75.27 / 75.39
(2)	27.05 / 37.02 / 31.26	24.24 / 43.98 / 31.25	24.4 / 56.88 / 34.15
(3)	36.47 / 13.19 / 19.37	33.08 / 31.82 / 32.43	40.97 / 43.7 / 42.29
(4)	40.31 / 54.08 / 46.19	**74.93** / 74.31 / **74.62**	77.19 / 79.52 / 78.34
(5)	35.24 / 12.3 / 18.24	43.7 / 47.28 / 45.42	35.35 / 42.39 / 38.55
(6)	56.16 / **67.58** / 61.34	67.99 / **77.08** / 72.25	**78.38** / **80.49** / **79.42**
(7)	78.79 / **71.82** / **75.14**	86.26 / **75.34** / 80.43	81.59 / **77.3** / 79.39
(8)	**92.9** / 61.53 / 74.03	**93.55** / 72.04 / 81.4	**92.58** / 73.93 / **82.21**
(9)	81.7 / 69.44 / 75.07	89.22 / 75.25 / **81.64**	85.76 / 75.27 / 80.17

Table 3. Study of F score per urban element. Column P indicates the pipeline that generated the best F score. If the pipeline or F score calculated with the hidden ground-truth differs from those ones calculated with the complete one, they are shown in square brackets. Pipelines are numbered as: (1) COLMAP + COLMAP, (2) COLMAP + OpenMVS, (3) OpenMVG-g + COLMAP, (4) OpenMVG-g + OpenMVS, (5) OpenMVG-i + COLMAP, (6) OpenMVG-i + OpenMVS, (7) Metashape, (8) Pix4D, (9) RealityCapture.

	AREA 1					
	oblique		nadir		combined	
	P	F score	P	F score	P	F score
facade	(7)	73.69 [72.25]	(7)	78.35 [76.69]	(7)	79.34 [77.82]
window	(7)	73.49 [71.72]	(7)	74.71 [73.67]	(7)	75.95 [74.46]
door	(7)	62.81 [62.46]	(7)	65.37 [64.32]	(7)	66.11 [64.42]
roof	(7)	88.66 [85.28]	(9)	91.04 [89.18]	(9) [(7)]	91.32 [89.53]
r. window	(7)	80.4 [79.9]	(9)	85.3 [85.77]	(8)	88.29 [86.8]
r. door	(8) [(4)]	60.84 [65.94]	(8)	72.46 [82.12]	(8)	82.44 [89.4]
sidewalk	(7)	83.55 [83.17]	(8)	90.64 [90.45]	(9)	90.15 [90.12]
street	(7)	85.76 [84.76]	(9)	92.73 [92.58]	(9)	92.46 [92.34]
grass	(9)	88.97 [87.96]	(9)	89.16 [87.92]	(9)	90.84 [89.98]
tree	(6)	31.74 [32.55]	(4)	38.47 [38.37]	(1) [(6)]	40.92 [40.21]
	AREA 2					
	oblique		nadir		combined	
	P	F score	P	F score	P	F score
facade	(7)	73.87 [71.99]	(7)	82.4 [81.67]	(7)	80.54 [79.71]
window	(8)	68.84 [69.89]	(7)	73.98 [74.1]	(7)	73.33 [73.1]
door	(7)	55.52 [57.13]	(7)	64.7 [64.87]	(7)	62.28 [63.26]
roof	(8)	90.08 [87.61]	(8)	93.34 [92.49]	(8)	94.26 [93.84]
r. window	(8)	88.35 [87.65]	(8)	88.55 [88.63]	(8)	89.25 [90.08]
r. door	(8)	72.64 [68.00]	(9) [(7)]	75.11 [67.43]	(9) [(8)]	74.04 [65.84]
sidewalk	(9)	87.59 [88.14]	(9)	91.92 [91.39]	(9)	91.99 [91.77]
street	(7)	86.04 [85.04]	(9)	92.55 [92.15]	(9)	92.83 [92.38]
grass	(9)	95.31 [93.43]	(8)	95.9 [93.79]	(6)	96.36 [94.91]
tree	(1)	25.34 [28.09]	(1)	33.7 [35.08]	(1)	39.45 [41.34]

per measurement. F score in Table 3, Precision in Table 4 and Recall in Table 5. Each row has the results of a specific class (i.e., urban category) and each column corresponds to a unique set of images. The result presented is the maximum score obtained among the nine pipelines tested (see Sect. 4.2) and the pipeline that generated the score is shown in column P. The results for the hidden area are presented in squared brackets if they differ from the ones calculated with the complete area.

Table 4. Study of precision per urban element. Column P indicates the pipeline that generated the best precision. If the pipeline or precision calculated with the hidden ground-truth differs from those ones calculated with the complete one, they are shown in square brackets. Pipelines are numbered as: (1) COLMAP + COLMAP, (2) COLMAP + OpenMVS, (3) OpenMVG-g + COLMAP, (4) OpenMVG-g + OpenMVS, (5) OpenMVG-i + COLMAP, (6) OpenMVG-i + OpenMVS, (7) Metashape, (8) Pix4D, (9) RealityCapture.

	AREA 1					
	oblique		nadir		combined	
	P	precision	P	precision	P	precision
facade	(7)	82.25 [81.35]	(7)	86.7 [85.94]	(7)	84.82 [83.82]
window	(7)	67.51 [67.01]	(9)	70.91 [70.56]	(7)	68.81 [68.24]
door	(8)	73.71 [74.08]	(9) [(8)]	67.02 [65.42]	(9)	66.31 [64.65]
roof	(9)	91.38 [90.61]	(9)	92.85 [92.21]	(9)	93.27 [92.61]
r. window	(8)	83.22 [89.29]	(8)	86.72 [87.92]	(8)	89.93 [89.62]
r. door	(8)	84.91 [85.2]	(9) [(8)]	77.16 [86.11]	(8)	78.24 [85.37]
sidewalk	(8)	97.63 [97.5]	(8)	97.12 [97.45]	(8)	96.13 [95.94]
street	(8)	97.31 [97.11]	(8)	97.37 [97.27]	(8) [(9)]	95.76 [95.62]
grass	(9)	95.21 [94.66]	(8)	98.03 [97.77]	(8)	96.95 [96.64]
tree	(9)	80.27 [79.5]	(8)	87.34 [86.8]	(8)	86.11 [85.44]
	AREA 2					
	oblique		nadir		combined	
	P	precision	P	precision	P	precision
facade	(8)	90.4 [89.8]	(7)	88.98 [88.93]	(8)	86.67 [85.76]
window	(8)	76.11 [74.17]	(9)	74.22 [73.14]	(8)	71.73 [69.62]
door	(8)	78.59 [76.08]	(8)	78.18 [77.27]	(8)	76.79 [75.53]
roof	(8)	97.18 [96.82]	(8)	96.34 [96.31]	(8)	96.38 [96.08]
r. window	(8)	95.51 [95.8]	(8)	93.45 [93.46]	(8)	93.83 [93.63]
r. door	(8)	79.94 [70.23]	(9) [(7)]	82.08 [78.08]	(8)	78.19 [69.25]
sidewalk	(8)	96.34 [96.11]	(8)	96.72 [96.31]	(8)	97.03 [96.74]
street	(8)	97.09 [97.02]	(8)	97.26 [96.98]	(8)	97.72 [97.52]
grass	(8)	99.48 [99.14]	(8)	99.5 [99.25]	(8)	99.43 [99.15]
tree	(8)	94.05 [93.58]	(8)	93.4 [92.97]	(8)	92.63 [92.05]

Analyzing the results that were obtained per class across all the image sets available, we can observe that although in [15] the method that most frequently got the maximum precision was COLMAP + COLMAP, the commercial applications are better in all the categories, and Pix4D is the one that most frequently gets the maximum precision. Roof, sidewalk, street and grass are the categories which obtained the best results. When looking at the recall, the pipeline OpenMVG-i + OpenMVS was the one that more frequently achieved the highest

Table 5. Study of recall per urban element. Column P indicates the pipeline that generated the best recall. If the pipeline or recall calculated with the hidden ground-truth differs from those ones calculated with the complete one, they are shown in square brackets. Pipelines are numbered as: (1) COLMAP + COLMAP, (2) COLMAP + OpenMVS, (3) OpenMVG-g + COLMAP, (4) OpenMVG-g + OpenMVS, (5) OpenMVG-i + COLMAP, (6) OpenMVG-i + OpenMVS, (7) Metashape, (8) Pix4D, (9) RealityCapture.

	AREA 1					
	oblique		nadir		combined	
	P	recall	P	recall	P	recall
facade	(7)	66.73 [64.98]	(7)	71.46 [69.23]	(7)	74.52 [72.62]
window	(7)	80.63 [77.14]	(7)	81.98 [78.85]	(7)	84.73 [81.93]
door	(7)	62.75 [63.33]	(4)	65.95 [66.12]	(4)	68.76 [67.99]
roof	(7)	87.53 [83.27]	(7)	90.34 [88.33]	(4)	92.25 [91.65]
r. window	(7) [(6)]	83.34 [84.01]	(7)	86.05 [86.28]	(4)	91.54 [93.07]
r. door	(4)	56.41 [71.34]	(8)	71.6 [78.49]	(8)	87.13 [93.84]
sidewalk	(7)	81.44 [80.82]	(4)	90.37 [91.4]	(4)	92.88 [92.56]
street	(7)	82.34 [81.48]	(4)	92.97 [92.75]	(4)	95.45 [95.68]
grass	(9)	83.49 [82.15]	(9)	83.95 [82.22]	(6)	88.45 [87.43]
tree	(6)	20.94 [21.65]	(4) [(6)]	25.82 [25.8]	(2)	29.44 [28.76]
	AREA 2					
	oblique		nadir		combined	
	P	recall	P	recall	P	recall
facade	(7)	70.17 [67.85]	(7)	76.73 [75.5]	(7)	78.1 [77.00]
window	(7)	70.24 [71.28]	(7)	74.93 [75.22]	(6)	77.05 [77.71]
door	(7)	52.32 [55.95]	(7)	62.84 [64.51]	(6)	68.94 [68.34]
roof	(7)	86.48 [84.19]	(8)	90.52 [88.96]	(8)	92.24 [91.71]
r. window	(8)	82.19 [80.78]	(8) [(6)]	84.14 [84.97]	(4) [(6)]	85.78 [87.82]
r. door	(8)	66.56 [65.91]	(9)	69.23 [60.44]	(9) [(8)]	71.81 [62.76]
sidewalk	(6)	84.66 [84.74]	(6)	93.77 [93.31]	(6)	94.71 [94.59]
street	(7)	83.1 [81.74]	(4)	95.47 [95.05]	(6)	96.31 [95.66]
grass	(7)	93.93 [91.61]	(4)	95.08 [92.82]	(6)	96.88 [95.67]
tree	(1)	15.03 [17.01]	(1)	21.62 [22.82]	(2)	26.58 [30.28]

scores among the open-source methods [15]. When we include the commercial software in the analysis, we see that they are not the best in all the image categories and sets of images, as it was the case with the precision. Now, on the results with the reconstruction from the combined set, we can see how the OpenMVG-g + OpenMVS and OpenMVG-i + OpenMVS are still the best in the majority of classes (in *Area 1* and n *Area 2*, respectively), in accordance with the results obtained in the scene level evaluation.

When looking at the F score, the class with lowest F score values is tree and the results are really influenced by the low values of the recall. We can see that for this class the winning methods are the open-source solutions whereas for the rest of the categories the commercial solutions are the best. These results confirm the hypothesis in [30]: trees in the parks of the city can degrade the scores of the reconstructions. We can also analyze the results depending on the image set under study. For example, with the combined set, according to [15] the pipelines with best performance in the majority of classes among the open-source techniques were OpenMVG-g + OpenMVS in *Area 1* and OpenMVG-i + OpenMVS in *Area 2*. These results are in accordance with the ones commented before, which does not consider the class information (Table 2). However, when looking at the nine pipelines, the methods RealityCapture and Pix4D (which were the best in Area 1 and Area 2, respectively, in the scene evaluation) have a different behaviour when looking at the specific urban elements. RealityCapture is the most frequent winner among all the urban categories, whereas Pix4D is only the winner in the roof and r. window (with around 31% of occupancy), suggesting that the main reason why this was the winner method is because it is the best in this specific category.

5.3 General Pipeline Evaluation

We can also observe that, in general, the open-source pipelines that obtained the best results are COLMAP + COLMAP, OpenMVG-g + OpenMVS and OpenMVG-i + OpenMVS. These results are in accordance with previous studies that used the same kind of metric, where COLMAP + COLMAP and OpenMVG-i + OpenMVS obtained the best results [7]. In particular, in that study OpenMVG-g + OpenMVS never has better results than COLMAP + COLMAP, but this situation is plausible in our study given the different camera trajectories (aerial grid configuration vs circle around an object), software versions and parameters used. Pix4D was also compared in that study, obtaining the best results for some categories, in accordance to what we observe in this study.

COLMAP used as MVS is better than OpenMVS only if it is applied after COLMAP SfM, whereas OpenMVS is better using the other SfM methods tested. This leads to the necessity to test not only a particular MVS method but a complete pipeline since it is going to be influenced by: the results obtained in the SfM step, the data conversion and preparation for the MVS step, as well as memory and computing limitations. The commercial applications used in the evaluation give better results than the open-source pipelines in general, and are prepared to handle large amounts of data, as part of their photogrammetry pipelines. Also, some of them provide by default a 3D mesh along with the point cloud. However, some have other disadvantages such as the limitation of supporting only certain operating systems, and all of them are not freely available.

6 Conclusion

We have presented in this paper a study of image-based 3D reconstruction pipelines in a metropolitan area, using exclusively aerial images. This study not only takes under consideration open-source 3D reconstruction techniques but also commercial photogrammetry solutions which are widely used in the industry. The final dense point cloud is evaluated at scene level and per urban category, thus allowing for a finer examination. Thanks to that, we see the influence of urban categories (e.g., roof) in the F scores. We also support the hypothesis done about how parks can degrade the F score values in a scene level evaluation (mainly because of the presence of trees) with facts. We have concluded that the commercial applications have better scores in the majority of the scenarios but the best solution of the open-source techniques is not far from them. When choosing the best 3D reconstruction pipeline other aspects apart from the F score might be important: the budget, the equipment, the possibility of adapting the algorithms for some specific needs, etc. In this study we have created an exhaustive and comprehensive review and we believe that it can be useful for those who want to see how 3D reconstruction methods perform using aerial images from a city.

Acknowledgements. This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under the Grant Number 15/RP/2776.

References

1. Aanæs, H., Jensen, R.R., Vogiatzis, G., Tola, E., Dahl, A.B.: Large-scale data for multiple-view stereopsis. Int. J. Comput. Vis. **120**(2), 153–168 (2016). https://doi.org/10.1007/s11263-016-0902-9
2. Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D.B.: Patchmatch: a randomized correspondence algorithm for structural image editing. ACM Trans. Graph. (ToG) **28**(3), 24 (2009)
3. Can, G., Mantegazza, D., Abbate, G., Chappuis, S., Giusti, A.: Semantic segmentation on swiss3dcities: a benchmark study on aerial photogrammetric 3d pointcloud dataset. Pattern Recogn. Lett. **150**, 108–114 (2021)
4. Furukawa, Y., Ponce, J.: Accurate, dense, and robust multiview stereopsis. IEEE Trans. Pattern Anal. Mach. Intell. **32**(8), 1362–1376 (2010)
5. Hartley, R., Zisserman, A.: Multiple view Geometry in Computer Vision. Cambridge University Press, Cambridge (2003)
6. Jancosek, M., Pajdla, T.: Multi-view reconstruction preserving weakly-supported surfaces. In: CVPR 2011, IEEE (2011)
7. Knapitsch, A., Park, J., Zhou, Q.Y., Koltun, V.: Tanks and temples: benchmarking large-scale scene reconstruction. ACM Trans. Graph. **36**(4), 1–13 (2017)
8. Menze, M., Geiger, A.: Object scene flow for autonomous vehicles. In: CVPR 2015 (2015)
9. Moulon, P., Monasse, P., Marlet, R.: Adaptive structure from motion with *a contrario* model estimation. In: Lee, K.M., Matsushita, Y., Rehg, J.M., Hu, Z. (eds.) ACCV 2012. LNCS, vol. 7727, pp. 257–270. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-37447-0_20

10. Moulon, P., Monasse, P., Marlet, R.: Global fusion of relative motions for robust, accurate and scalable structure from motion. In: ICCV 2013 (2013)
11. Munoz, D., Bagnell, J.A., Vandapel, N., Hebert, M.: Contextual classification with functional max-margin markov networks. In: CVPR 2009 (2009)
12. Rottensteiner, F., Sohn, G., Gerke, M., Wegner, J.D., Breitkopf, U., Jung, J.: Results of the ISPRS benchmark on urban object detection and 3d building reconstruction. ISPRS J. Photogrammetry Remote Sens. **93**, 256–271 (2014)
13. Ruano, S., Cuevas, C., Gallego, G., García, N.: Augmented reality tool for the situational awareness improvement of UAV operators. Sensors **17**(2), 297 (2017)
14. Ruano, S., Gallego, G., Cuevas, C., García, N.: Aerial video georegistration using terrain models from dense and coherent stereo matching. In: Geospatial InfoFusion and Video Analytics IV; and Motion Imagery for ISR and Situational Awareness II. International Society for Optics and Photonics (2014)
15. Ruano, S., Smolic, A.: A benchmark for 3d reconstruction from aerial imagery in an urban environment. In: VISIGRAPP (5: VISAPP), pp. 732–741 (2021)
16. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: CVPR 2016 (2016)
17. Schönberger, J.L., Zheng, E., Frahm, J.-M., Pollefeys, M.: Pixelwise view selection for unstructured multi-view stereo. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9907, pp. 501–518. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46487-9_31
18. Schops, T., et al.: A multi-view stereo benchmark with high-resolution images and multi-camera videos. In: CVPR 2017 (2017)
19. Seitz, S.M., Curless, B., Diebel, J., Scharstein, D., Szeliski, R.: A comparison and evaluation of multi-view stereo reconstruction algorithms. In: CVPR 2006 (2006)
20. Serna, A., Marcotegui, B., Goulette, F., Deschaud, J.E.: Paris-rue-madame database: a 3d mobile laser scanner dataset for benchmarking urban detection, segmentation and classification methods (2014)
21. Shahid, M., et al.: Aerial cross-platform path planning dataset. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3936–3945 (2021)
22. Snavely, N., Seitz, S.M., Szeliski, R.: Modeling the world from internet photo collections. Int. J. Comput. Vis. **80**, 189–210 (2008)
23. Stathopoulou, E.K., Welponer, M., Remondino, F.: Open-source image-based 3d reconstruction pipelines: review, comparison and evaluation. The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences (2019)
24. Strecha, C., Von Hansen, W., Van Gool, L., Fua, P., Thoennessen, U.: On benchmarking camera calibration and multi-view stereo for high resolution imagery. In: CVPR 2008 (2008)
25. Sweeney, C., Hollerer, T., Turk, M.: Theia: A fast and scalable structure-from-motion library. In: Proceedings of the ACM International Conference on Multimedia, pp. 693–696. ACM (2015)
26. Vallet, B., Brédif, M., Serna, A., Marcotegui, B., Paparoditis, N.: Terramobilita/iqmulus urban point cloud analysis benchmark. Comput. Graph. **49**, 126–133 (2015)
27. Yan, F., Xia, E., Li, Z., Zhou, Z.: Sampling-based path planning for high-quality aerial 3d reconstruction of urban scenes. Remote Sens. **13**(5), 989 (2021)
28. Zhang, L., Li, Z., Li, A., Liu, F.: Large-scale urban point cloud labeling and reconstruction. ISPRS J. Photogrammetry Remote Sens. **138**, 86–100 (2018)

29. Zhou, W., Liu, J., Lei, J., Yu, L., Hwang, J.N.: Gmnet: graded-feature multilabel-learning network for RGB-thermal urban scene semantic segmentation. IEEE Trans. Image Process. **30**, 7790–7802 (2021)
30. Zolanvari, S., et al.: DublinCity: annotated lidar point cloud and its applications. In: 30th BMVC (2019)