Tanks and Temples: Benchmarking Large-Scale Scene Reconstruction

ARNO KNAPITSCH, JAESIK PARK, QIAN-YI ZHOU, and VLADLEN KOLTUN, Intel Labs

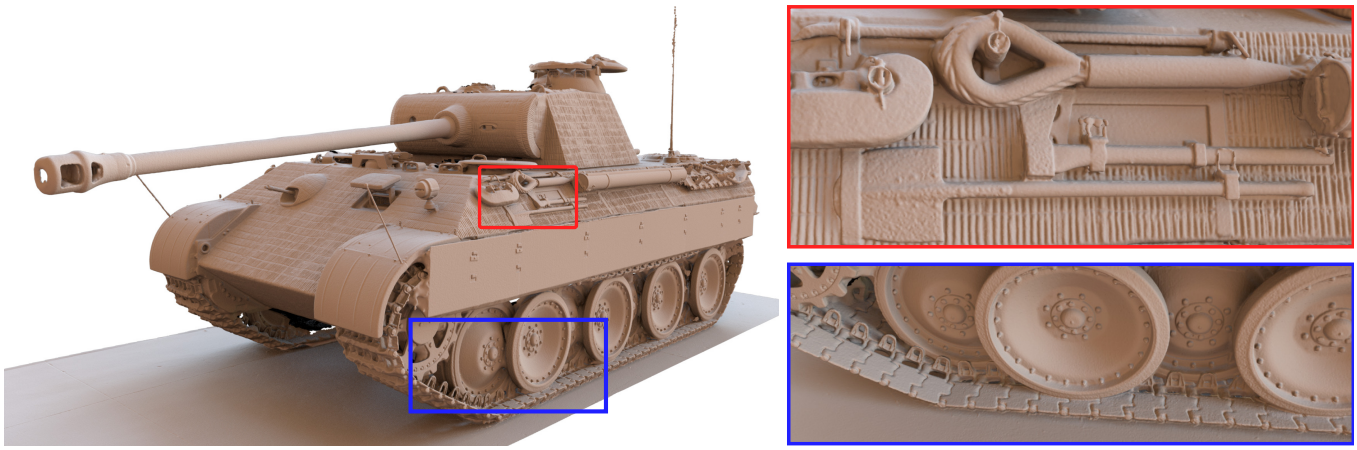

Fig. 1. Ground-truth model for the Panther dataset, one of the datasets in the presented benchmark for large-scale scene reconstruction.

ABSTRACT

We present a benchmark for image-based 3D reconstruction. The benchmark sequences were acquired outside the lab, in realistic conditions. Ground-truth data was captured using an industrial laser scanner. The benchmark includes both outdoor scenes and indoor environments. High-resolution video sequences are provided as input, supporting the development of novel pipelines that take advantage of video input to increase reconstruction fidelity. We report the performance of many image-based 3D reconstruction pipelines on the new benchmark. The results point to exciting challenges and opportunities for future work.

Additional Key Words and Phrases: Structure from motion, multi-view stereo, image-based reconstruction, large-scale scene reconstruction

This is the authors' preprint of a SIGGRAPH 2017 paper.

1 INTRODUCTION

In the past decade, structure from motion (SfM) and multi-view stereo (MVS) techniques have advanced to enable remarkable reconstructions of landmark scenes from community photo collections [Agarwal et al. 2011; Shan et al. 2013; Snavely et al. 2008]. Due to these accomplishments, image-based reconstruction is deservedly considered one of the great successes of visual computing, combining rigorous theory [Hartley and Zisserman 2000; Triggs et al. 2000], advanced computational methods [Agarwal et al. 2010; Furukawa and Hernández 2015; Wu et al. 2011], and a culture of open software development [Fuhrmann et al. 2015; Furukawa 2011; Moulon et al. 2016; Schönberger 2016; Snavely 2010; Wu 2011].

Nonetheless, existing reconstruction techniques have significant limitations. Take a nearby camera, walk around a building while recording a video, and feed the data into a standard SfM+MVS pipeline. There is a good chance that the result will not be a clean and accurate reconstruction of the building. For an even greater challenge, take a video while walking around the interior of your residence and use that. A clean and complete 3D reconstruction of the environment is unlikely to emerge.

How can these limitations coexist with the remarkable successes of image-based reconstruction? Reconstruction from community photo collections involves dealing with massive and highly redundant datasets. Processing such datasets brings up significant challenges in system and algorithm design [Agarwal et al. 2011; Frahm et al. 2010; Heinly et al. 2015; Schönberger and Frahm 2016; Wu 2013]. But it also affords significant freedom in automatically selecting a subset of the input data that can be successfully reconstructed [Furukawa et al. 2010; Goesele et al. 2007; Li et al. 2008; Schönberger et al. 2016]. Such data selection enables challenging input to be discarded and can leave limitations in the underlying

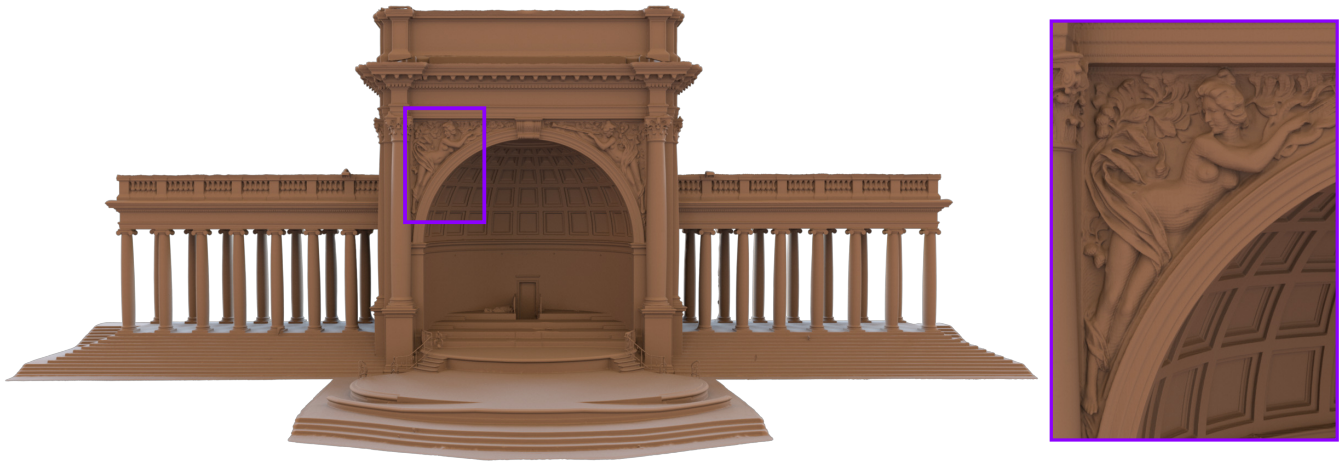

Fig. 2. Ground-truth model for the Temple dataset. This scene has an area of 713 square meters and a height of 21 meters. The point sets for this and other datasets were meshed to create the renderings shown in the paper. As a result, the renderings may exhibit meshing artifacts that are not present in the ground-truth point sets.

techniques unaddressed. Furthermore, the most spectacular results of these pipelines are not evaluated against precise ground-truth models, since such models are not available. This can discourage further improvement in accuracy and completeness, since even if real improvement is made, it is hard to quantitatively substantiate.

The most commonly used multi-view stereo benchmarks focus on small tabletop objects imaged in laboratory conditions [Aanæs et al. 2016; Seitz et al. 2006]. They provide ground-truth poses for a uniformly distributed set of outside-in views as input, thus restricting attention to one part of the 3D reconstruction pipeline and eliminating difficulties that cause failure in large-scale real-world environments. The benchmarks of Strecha et al. [2008] and Merrell et al. [2007] attempted to address some of these issues, but have a number of limitations. As a result, they also do not push existing pipelines beyond their limits.

In order to stimulate progress on some of the standing challenges in large-scale scene reconstruction, we have created a new benchmark. To acquire ground-truth models of large-scale scenes, we used a state-of-the-art industrial laser scanner with a range of 330 meters and submillimeter accuracy. The scanner can acquire up to a million points per second. We have scanned objects and environments from multiple viewpoints and registered the scans to obtain ground-truth models. For each model, we provide 8-megapixel video as input for reconstruction.

The presented benchmark has a number of characteristics that can support the development of new reconstruction techniques:

- The input modality is video. This can help future pipelines track the camera, reason about illumination and reflectance, and reconstruct small details.
- The benchmark evaluates complete reconstruction pipelines. This leaves scope for tackling camera localization and dense reconstruction jointly, potentially increasing robustness and precision via co-adaptation to the performance characteristics of each task.
- The benchmark includes both outdoor and indoor scans of complete scenes, pushing current reconstruction pipelines to their limits and beyond.

The presented datasets are organized into two groups: intermediate and advanced. The intermediate group contains sculptures, large vehicles, and house-scale buildings with outside-looking-in camera trajectories. The advanced group contains large indoor scenes imaged from within and large outdoor scenes with complex geometric layouts and camera trajectories.

We have evaluated many SfM+MVS pipelines on the presented benchmark. The results indicate that image-based 3D reconstruction is far from solved. Existing pipelines perform impressively given the difficulties, but the need for significant progress is clear.

2 EXISTING BENCHMARKS

The Middlebury benchmark of Seitz et al. [2006] put multi-view stereo research on a quantitative footing and was instrumental in directing efforts in the area. The benchmark is based on two small objects, 10-20 cm across, with nearly Lambertian surfaces. Ground-truth models were acquired using a laser stripe scanner. Input images were acquired by a VGA-resolution camera mounted on a precisely controlled robotic arm. The objects were imaged in a lab with controlled lighting. The object is seen from regularly spaced positions around it. Accurate camera poses for every image are provided as input. Due to these simplifying factors, methods evaluated by the authors performed remarkably well even at the time, consistently achieving submillimeter accuracy.

Aanæs et al. [2016] constructed a larger MVS benchmark using 80 tabletop arrangements. The objects are deliberately more challenging and have a variety of materials, including specular ones. The image resolution is also higher: 1600×1200. Other key characteristics

of the Middlebury benchmark were retained: small tabletop object arrangements were imaged in the lab, under controlled lighting, by a camera positioned at regularly spaced viewpoints using a precisely controlled robotic arm. Ground-truth camera poses are given as input. In comparison, our work concerns large-scale outdoor and indoor scenes, imaged in realistic conditions, with high-resolution video input.

The EPFL benchmark of Strecha et al. [2008] used building facades and was acquired outside the lab. The input images have high resolution (6.2 MP). Dense ground-truth models were acquired by a LiDAR scanner. This benchmark provided a significant challenge and supported the development and validation of advanced reconstruction pipelines [Langguth et al. 2016; Schönberger et al. 2016; Tola et al. 2012; Vu et al. 2012]. It is an important precursor to our work. Nevertheless, it has a limited scope: only three building facades are included.

The UNC dataset of Merrell et al. [2007] is perhaps closest to ours in motivation: they focus on large-scale scene reconstruction, advocate video as the input form, and specifically target realistic acquisition conditions. ("The data is representative of the kind of data expected in a real-world application where there are many uncontrolled variables such as variations in texture, brightness, and the distance between the camera and the scene.") On the other hand, this dataset has significant limitations. It contains only a single scene, a building, and the reference model has limited fidelity. The reference model consists of a small number of planar facets produced by a surveying procedure, and does not represent the geometric details of the underlying building.

Our benchmark combines and extends the most compelling characteristics of the EPFL and UNC datasets: video input (UNC), dense and precise ground-truth (EPFL), and high-resolution input (EPFL). In fact, our datasets have the highest input resolution, at video rate, and the most precise ground truth. Crucially, our benchmark is much larger and more diverse, incorporating complete large-scale outdoor structures as well as complex indoor environments.

Concurrently with our work, Schöps et al. [2017] created a new benchmark for two- and multi-view stereo algorithms. Their benchmark provides input images at very high resolution (24 MP), as well as image sequences captured with arrays of synchronized low-resolution cameras (0.4 MP). While our benchmark evaluates full scene reconstruction pipelines, theirs focuses on evaluating binocular stereo and MVS. Thus the two benchmarks are complementary.

When ground-truth geometry is not available, 3D reconstructions can still be evaluated by measuring the realism of rendered images, either via perceptual experiments with human observers [Choi et al. 2015; Shan et al. 2013] or by automatically comparing to corresponding real images [Waechter et al. 2017].

There are also benchmarks for related problems, such as mesh reconstruction from point clouds [Berger et al. 2013], visual odometry [Burri et al. 2016; Geiger et al. 2013], and RGB-D reconstruction [Choi et al. 2015; Handa et al. 2014; Sturm et al. 2012]. Our work deals with a different problem: scene reconstruction from images or video.

3 KEY DECISIONS

Video. One of the distinguishing characteristics of the presented benchmark is the focus on video as the input modality. This is not a trivial choice. While a number of projects have considered 3D reconstruction from video [Frahm et al. 2010; Kolev et al. 2014; Newcombe et al. 2011; Pollefeys et al. 2008; Schöps et al. 2015; Tanskanen et al. 2013; Vogiatzis and Hernández 2011; Wendel et al. 2012], much more work in the literature is devoted to reconstruction from image collections. This is in part due to the low resolution and image quality that characterized digital video cameras in the past. As an informal but representative datapoint, the camera on the iPhone 4, released in 2010, captured 5 MP images but only 0.9 MP video. In general, high-fidelity and high-resolution photographs were more accessible in the past than digital video with comparable fidelity and resolution.

This has changed in the last few years. The camera on the iPhone 6s, released in 2015, captures 8.3 MP video at 30 fps. (For comparison, the high-resolution EPFL benchmark used 6.2 MP images [Strecha et al. 2008].) The fidelity of digital video has also increased dramatically. On the high end, the Blackmagic Production Camera, which was one of the cameras acquired for this project, provides global-shutter 8.6 MP video with 12 stops of dynamic range. Digital video with high resolution, high framerate, high-quality optics and sensors, high dynamic range, and minimal rolling-shutter artifacts is increasingly available.

In fact, since resolution and image fidelity are not the bottlenecks they once were, video may become the default input modality for 3D reconstruction. This is due to multiple advantages that video has over image collections:

- Data redundancy can be leveraged to increase reconstruction fidelity. A complex surface will generally be imaged more thoroughly and with greater coverage by a continuously capturing video camera.
- Camera localization can benefit from the presence of both narrow and wide baselines. In particular, video can support localization in challenging scenes, for example in the presence of repetitive structures.
- Video can assist the estimation of reflectance properties, by providing more data about specular highlights and view-dependent surface appearance.
- Redundant surface sampling can be used for super-resolution, potentially increasing the accuracy of both geometry and estimated material properties.
- Video capture has significant usability advantages. The operator need only turn the camera on and move it through the environment, without also making decisions about the timing of individually captured frames. All frames along the camera trajectory are captured and can be used. The usability advantage becomes particularly notable in settings in which it is difficult to explicitly supervise and trigger the camera during operation because the operator's attention is consumed by other tasks, such as piloting the UAV or steering the vehicle on which the camera is mounted.
- Video is the default imaging modality for an increasing range of devices, such as UAVs. Cameras for such devices

Fig. 3. Other models from the intermediate group.

are increasingly selected and manufactured to optimize video capture, rather than the acquisition of still images.

These advantages coexist with challenges brought up by video, such as high data volumes and processing demands. We decided to collect video input to support investigation of the advantages as well as the challenges. Ultimately, given the high resolution of the video we provide, the sequences can always be temporally subsampled and treated as image collections if desired, whereas the opposite transformation is hardly possible.

Outdoor and indoor. Another key characteristic of the presented benchmark is that it includes both outdoor and indoor scenes. While

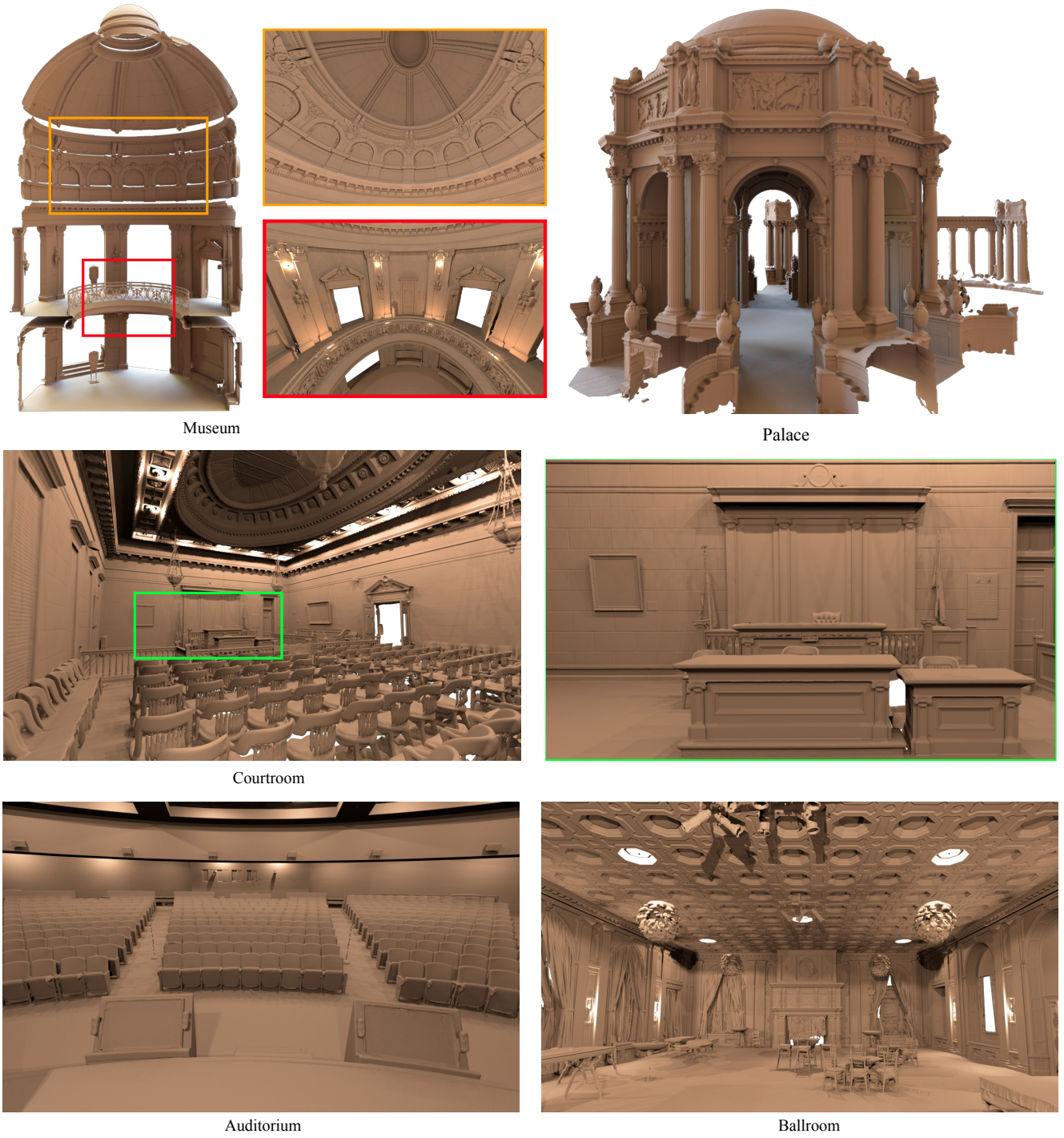

Fig. 4. Other models from the advanced group.

image-based reconstruction of outdoor environments is considered fairly well-understood, high-fidelity reconstruction of indoor scenes from images or video is known to be extremely challenging and is generally regarded as an open problem. The most prominent results to date use structure priors [Furukawa et al. 2009; Ikehata et al. 2015; Xiao and Furukawa 2014] or depth cameras [Choi et al. 2015; Zhou and Koltun 2013].

Since indoor environments provide the setting for so much of our lives, we believe that high-fidelity reconstruction of such environments should be considered a core requirement for modern pipelines. Indoor environments are not governed by wholly distinct geometric and physical laws: the same fundamentals apply indoors and out. We provide multiple comprehensive scans of large-scale indoor environments. Our goal is to accelerate progress towards reliable broad-competence systems.

Complete pipeline. Rather than focusing on subproblems, such as SfM or MVS in isolation, the presented benchmark was created to evaluate complete 3D reconstruction pipelines: from input video to dense point cloud. This can support novel approaches that reconsider the interface between camera localization and dense reconstruction. Direct methods have demonstrated that the camera can be localized and a dense or semi-dense 3D model can be created without sparse feature matching [Engel et al. 2017, 2014; Newcombe et al. 2011]. While such methods have not displaced the traditional SfM+MVS pipelines for high-fidelity scene reconstruction, we want to support experimentation with novel formulations that tackle the reconstruction problem as a whole and allow parts of the system to co-adapt.

High-end camera. Given our motivation to support the development of robust scene reconstruction pipelines that can be widely used, and given the rapid improvement of camera modules for mobile devices, we could have used a smartphone camera to capture input video for the benchmark. This would have guaranteed that high reconstruction fidelity on the benchmark immediately translates to videos that millions of people can capture with devices already in their possession.

We decided against this because mobile cameras continue to improve rapidly and we do not want to saddle the benchmark with optical aberrations that may be largely irrelevant in a few years, or already are when a dedicated camera is used. Features found in high-end cameras tend to make their way into lower-end sensors and ISPs. We thus use a high-end videography setup with fast professional lenses, high light sensitivity in indoor environments, wide field of view, and gimbal stabilization.

One consequence of this decision is that the benchmark is backward compatible. Existing SfM+MVS pipelines that were designed for image collections can be applied to collections of frames from our videos, because the basic assumptions made by these pipelines are still applicable. This provides a path for gradual improvement starting from existing systems.

4 DATA COLLECTION

4.1 Ground truth

Scanning. Ground-truth data was collected using a FARO Focus 3D X330 HDR scanner. This laser scanner has a range of 330 meters and can operate both indoors and under direct sunlight. The scanner has a horizontal range of 360° and a vertical range of 300°. (There is a “blind cone" with an angular diameter of 60° centered at the vertical ray extending directly downwards from the scanner.) At the time of calibration (shortly before data acquisition), the ranging noise of the scanner was 0.1 mm at a distance of 10.2 meters and 0.3 mm at a distance of 22.7 meters.

The scanner can capture up to 976,000 points per second. Lateral sampling density can be traded off against capture time. An omnidirectional scan (360° horizontal, 300° vertical) at full resolution takes two hours and yields a lateral spacing of 0.3 mm at a distance of 2 meters. We usually operated the scanner at half or quarter resolution, which increases lateral spacing proportionately but does not diminish ranging accuracy. In most scenes, the horizontal range was restricted to a subset of 360° that covered the relevant part of the scene (e.g., a building or a vehicle). In indoor environments, full omnidirectional scans were taken since the entire surroundings were relevant.

In every scene, multiple scans had to be acquired to densely cover the surfaces. Since the scans have to be registered to each other, they must overlap. Simple objects such as small statues were scanned from 4 positions. Mid-sized structures such as the train were scanned from 8 to 10 positions. The biggest outdoor scenes – Palace and Temple – were scanned from 14 and 17 positions, respectively. Data acquisition for the Palace dataset spanned two days. Indoor environments with complex layouts also required many scans.

Artifacts and outliers. Even with careful scanner operation, outliers cannot be avoided in certain settings. Since some of the scenes are prominent architectural structures and since scanning takes hours, people inevitably walk through the scene while it is being scanned. If crowding became too strong, the scan was terminated and later repeated. But measurements of people rather than the underlying scene are still sometimes present in the scans. Most such measurements are removed by the scanner’s built-in noise filtering. We examined the scans and verified that the remaining outliers constitute less than one thousandth of all measurements and do not materially affect the evaluation.

Bodies of water, glass, and mirrors can produce mirror image artifacts. If the reflective surface is planar, the artifacts are usually clearly clustered and can be removed manually. If the reflective surface has complex geometry (e.g., polished metal ornaments, glasses, bottles), the spurious points are harder to remove and are usually left in. These artifacts are again quantitatively negligible and do not materially affect the evaluation.

Post-processing. The individual scans are registered using software provided by FARO. The software performs global alignment between all scans and reports point-to-point error in the overlap regions after alignment. We could thus verify that the inter-scan alignment error is in line with intra-scan point spacing. After alignment, the scene is examined and cleaned of prominent outliers, and the resulting point cloud is cropped to the area of interest. The point cloud can be extremely large: e.g., more than 600 million points for the Palace. The density of the data also varies, due to overlap between scans and varying distance of the scene’s surfaces from the scanner. We therefore resample the point cloud using a uniform voxel grid. The voxel size is set to $\tau/2$, where τ is the distance threshold listed for each scene in Table 1. When multiple points fall into the same voxel, the mean of these points is retained.

The distance threshold τ cannot be the same for all scenes, since the scenes vary drastically in scale and sampling density. (The Palace

is nearly 50 meters tall and scanned from the ground, while the tanks and the smallest statues are scanned at close range.) We set τ for each scene by examining the data and computing statistics of nearest-neighbor distances in the ground-truth point clouds.

4.2 Input video

We have acquired and tested three video cameras. The first has a global shutter (GS): the Blackmagic Production Camera, used with a Rokinon 10mm f/2.8 lens. The other two cameras have rolling shutters (RS): the DJI Zenmuse X5R, used with an Olympus M.Zuiko 12mm f/2.0 lens, and the Sony a7S II, used with a Zeiss Loxia 21mm f/2.8 lens. All cameras capture 4K (>8 MP) video. The DJI X5R is stabilized by a DJI Osmo gimbal and the Sony a7S II is stabilized by a Pilotfly H2 gimbal.

In preparation for data acquisition for the benchmark, we have tested existing reconstruction pipelines with video sequences from RS and GS cameras. We found that the gimbal eliminates fast camera motion to the extent that rolling-shutter distortion on videos captured by the X5R and the Sony a7S II became insignificant. In particular, we captured corresponding RS and GS sequences for four different scenes and used these as input. Reference pipelines such as COLMAP [Schönberger 2016] never performed worse on the RS sequences than on the corresponding GS sequences. We therefore concluded that reconstruction error is dominated by factors other than rolling shutter distortion. Due to the better low-light sensitivity of the RS cameras, we used the gimbal-stabilized X5R and Sony a7S II to produce video for the benchmark.

In choosing a lens for each camera, we had to consider the benefits and drawbacks of large field of view (FOV). A lens with a large FOV helps camera tracking, but can have stronger distortion. Many SfM systems cannot handle the distortion of lenses with FOV above 100°, since they use a simple camera model with one or two radial distortion parameters. The 12mm lens we chose for the X5R combines a large FOV (84° diagonal) and a lens that conforms to the simple models used by most systems (as evidenced by reprojection error during camera calibration). For the Sony a7S II, the 21mm lens provides a diagonal FOV of 90°.

For each scene, we captured test videos and reviewed them on site before capturing sequences for the benchmark, to make sure that the camera's settings are appropriate for the scene's light levels. The white balance was set manually and fixed for each scene. We kept exposure time below 10 milliseconds to minimize motion blur. For most scenes, the lens was focused to the hyperfocal distance.

We kept as many camera settings fixed for each scene as possible. Some scenes had to be recorded with automatic ISO, aperture, or shutter speed, due to high dynamic range in the scene. This was the case for example for outdoor scenes that were filmed on sunny days, such that some surfaces are under direct sunlight and others are in the shade. The camera settings used for each scene are reported in Table 1.

5 SCENES

The benchmark datasets are summarized in Table 1. The ground-truth point clouds are visualized in Figures 1, 2, 3, and 4. We now describe each scene in more detail.

Name	Cam	Area (m^2)	Height (m)	τ (mm)	Frames	Points (M)	ISO	f	Shutter (sec.)
Intermediate									
Family	S	5	2.1	3	4,395	5.5	640	f/3.2	1/160
Francis	S	81	15.2	5	7,830	19.3	Auto	f/7.1	1600
Horse	S	10	3.2	3	6,015	6.2	640	f/3.2	1/160
Lighthouse	D	108	11.1	10	8,322	8.2	200	f/4.0	Auto
M60	D	35	3.2	5	5,616	9.7	400	f/2.0	1/100
Panther	D	34	2.9	5	6,570	12.3	400	f/2.0	1/100
Playground	D	54	2.8	10	7,463	1.7	200	f/2.8	Auto
Train	S	35	5.6	5	12,630	21.7	Auto	f/5.6	1/1000
Advanced									
Auditorium	S	541	6.2	10	14,640	53.4	Auto	f/2.8	1/125
Ballroom	S	254	3.9	10	10,800	43.9	6000	f/3.2	1/160
Courtroom	S	206	7.8	10	7,049	43.4	1600	Auto	1/100
Museum	S	110	21.2	10	17,115	36.5	Auto	f/3.2	1/200
Palace	D	4,295	47.2	30	21,871	41.9	Auto	f/3.2	Auto
Temple	S	713	20.7	15	17,475	33.4	Auto	f/5.6	1/640

Table 1. Benchmark datasets. From left to right: Camera model, Sony a7S II (S) or DJI X5R (D); footprint area of the cropped ground-truth point set; height of the cropped ground-truth point set; the threshold τ used for precision and recall computation; number of frames in the input video; number of points in the ground-truth point set after cropping and subsampling (millions). The last three columns summarize the exposure settings used for video capture: ISO, f-number, and shutter speed.

5.1 Intermediate datasets

Family. The statue was filmed right after sunset, which allowed us to fix all exposure settings. The illumination is almost uniform from all directions.

Francis. This is the biggest sculpture in the benchmark, extending to a height of 15 meters. Due the sculpture's height, the camera must frequently look up; in those frames, the only background is the sky. Much of the sculpture is also symmetric. These factors can complicate camera localization. In some frames, the sun is directly behind the sculpture.

Horse. This bronze statue is highly specular and rests on a platform with a uniform specular surface. It was imaged right after sunset. This dataset is particularly challenging due to the uniform specular materials.

Lighthouse. The Lighthouse is one of the tallest structures in the intermediate group. It is imaged with only the sky in the background for parts of the sequence. On the other hand, it has detailed texture due to weathering on the bricks, which can help tracking and reconstruction.

M60. The M60A1 Patton battle tank was manufactured in the 1970s and was deployed in the first Gulf War. It is located inside a hangar with one side open. Additional light was switched on inside the hangar to assist video capture. The tank's surface is nearly Lambertian and is richly textured due to age. All exposure settings were fixed.

Panther. The Panther Mark V tank was left behind in a swamp in Poland during WWII, then salvaged and refurbished after more

than 40 years using mostly original parts and materials. The surface is covered with Zimmerit coating, which adds high-frequency geometric detail.

Playground. This scene was imaged on an overcast day under fairly uniform illumination. The main difficulty is due to thin structures such as bars, poles, chains, and wooden beams. The reflective surface of the slide is an additional challenge.

Train. The train was imaged on a sunny day. The ISO was set to automatic to cope with varying illumination. Both foreground and background are sharp due to the small aperture.

5.2 Advanced datasets

The datasets in this group are challenging due to their scale, complexity, and other complicating factors. The four indoor scenes are in this group. The indoor datasets present a number of difficulties: illumination is considerably weaker, the camera may take in only a small part of the environment at a given time, and many surfaces are nearly uniform in appearance. The outdoor scenes in this group are challenging primarily due to their scale and complexity. All datasets were recorded with the best camera settings afforded by the scene, with the requirements of SfM and MVS techniques in mind.

Auditorium. This large auditorium was filmed with automatic ISO due to high dynamic range within the scene. Parts of the scene are strongly illuminated by spotlights while other parts receive only weak indirect illumination. This scene is challenging due to uniform regions, repeating structures, sharply varying illumination, and glare.

Ballroom. The ballroom has elaborate wooden paneling and a textured carpet. The room is populated by scattered tables and chairs. The dark interior and weak illumination necessitated a very high ISO setting. This is the only indoor scene filmed with fixed camera exposure settings.

Courtroom. This courtroom was built in 1910. A stained glass dome admits natural light. The scene is filled with regularly arranged wooden furniture. A patterned carpet and ornamentation add texture and structural detail.

Museum. This atrium is 21 meters high and was imaged from two stories. The stained glass dome admits a lot of natural light. The scene is additionally illuminated by artificial light sources. The aspect ratio is unusual: the scene is much higher than it is wide.

Palace. The Palace of Fine Arts in San Francisco is our largest scene in terms of both area and height. For this scene, we provide multiple video sequences concatenated together, rather than one continuous shot. This was necessary due to the layout of the physical environment, which necessitated imaging from several distributed locations. This scene was imaged with aperture priority and automatic shutter speed and ISO. This was necessary due to strongly varying illumination conditions in the scene, which contains areas under direct sunlight as well as shaded areas inside the dome. The main structure is highly symmetrical and contains many repeating patterns. The large body of water in front of the palace is an additional source of difficulty.

Temple. The Temple of Music in San Francisco was captured with automatic ISO settings on a slightly overcast day. The illumination varies strongly between directly illuminated and shaded areas. The Temple is large, but not as complex as the Palace.

5.3 Training datasets

We have collected additional datasets that can be used for training. The ground-truth models for these datasets will be made public. The datasets can be obtained from the benchmark's web site, www.tanksandtemples.org.

6 EVALUATION PROCEDURE

Since the evaluated reconstruction pipelines operate on image collections rather than video, we sampled a set of frames from each benchmark sequence and used these sets of frames to evaluate existing pipelines. For the reported evaluation, we sampled the video at regular intervals. We sampled 150 frames for Family and Horse, 500 for Palace, and 300 for all other scenes.

Alignment. For benchmarking, the point clouds produced by the evaluated pipelines must be aligned to the ground-truth models. Most pipelines expose the reconstructed camera poses, and for these we perform the alignment automatically. The reconstructed camera poses are registered to estimated ground-truth camera poses, yielding scale and pose estimates for the reconstructed point cloud. We estimate the ground-truth camera poses using the ground-truth point cloud [Mastin et al. 2009]. Alternatively, for pipelines that do not expose the camera poses (e.g., Pix4D), we manually align the reconstructed point cloud to the ground truth.

The approximate alignment described in the previous paragraph is used to initialize $Sim(3)$ refinement of the reconstruction to the ground-truth model. The approximate alignment provides a set of rough correspondences $\{(\mathbf{p}_i^m, \mathbf{q}_i^m)\}$. A refined pose and scale are estimated by optimizing the following objective:

$$E(\mathbf{T}) = \sum_i \|\mathbf{p}_i^m - \mathbf{T}\mathbf{q}_i^m\|^2. \tag{1}$$

Here the points are represented in homogeneous coordinates and $\mathbf{T}^m \in \mathbb{R}^{4\times4}$ is a similarity transformation:

$$\mathbf{T}^m = \begin{bmatrix} c\mathbf{R} & \mathbf{t} \\ \mathbf{0} & 1 \end{bmatrix}. \tag{2}$$

The least-squares estimate is obtained using Umeyama's algorithm [Umeyama 1991]. Starting from this scaling and alignment $\mathbf{T}^m$, we use an extension of ICP to similarity transformations (including scale) to refine the registration of the dense point clouds.

Resampling. The aligned reconstruction is resampled using the same voxel grid as the ground-truth point cloud. We again use voxel size $\tau/2$. When multiple points fall into the same voxel, the mean of these points is retained.

Cropping. Each ground-truth model is accompanied by a bounding volume, defined by a polygonal prism. The base polygon can have arbitrary complexity and was manually specified in an interactive interface. The bounding volume specifies the region in which reconstructions are evaluated against the ground-truth. The reconstructed point cloud is cropped to the interior of this bounding volume.

Measures. Let $\mathcal{G}$ be the ground truth and $\mathcal{R}$ a reconstructed point set being evaluated. For a reconstructed point $\mathbf{r} \in \mathcal{R}$, its distance to the ground truth is defined as

$$e_{\mathbf{r}\to\mathcal{G}} = \min_{\mathbf{g}\in\mathcal{G}} \|\mathbf{r}-\mathbf{g}\|. \tag{3}$$

These distances can be aggregated to define the *precision* of the reconstruction $\mathcal{R}$ for any distance threshold d:

$$P(d) = \frac{100}{|\mathcal{R}|} \sum_{\mathbf{r}\in\mathcal{R}} \left[e_{\mathbf{r}\to\mathcal{G}} < d\right], \tag{4}$$

where $[\cdot]$ is the Iverson bracket. $P(d)$ is defined to lie in the range [0,100] for convenience and can be interpreted as a percentage.

Similarly, for a ground-truth point $\mathbf{g} \in \mathcal{G}$, its distance to the reconstruction is defined as

$$e_{\mathbf{g}\to\mathcal{R}} = \min_{\mathbf{r}\in\mathcal{R}} \|\mathbf{g}-\mathbf{r}\|. \tag{5}$$

The *recall* of the reconstruction $\mathcal{R}$ for a distance threshold d is defined as

$$R(d) = \frac{100}{|\mathcal{G}|} \sum_{\mathbf{g}\in\mathcal{G}} \left[e_{\mathbf{g}\to\mathcal{R}} < d\right]. \tag{6}$$

Precision and recall can be combined in a summary measure, the *F-score*:

$$F(d) = \frac{2P(d)R(d)}{P(d)+R(d)}. \tag{7}$$

The F-score at a given threshold d is the harmonic mean of precision and recall at this threshold. It has the property that if either $P(d) \to 0$ or $R(d) \to 0$, then $F(d) \to 0$. It is thus a better summary measure than the arithmetic mean, which does not have this property.

The precision quantifies the accuracy of the reconstruction: how closely the reconstructed points lie to the ground truth. The recall quantifies the reconstruction's completeness: to what extent all the ground-truth points are covered. Precision alone can be maximized by producing a very sparse set of precisely localized landmarks. Recall alone can be maximized by densely covering the space with points. However, either of these schemes will drive the other measure and the F-score to 0. A high F-score for a stringent distance threshold can only be achieved by a reconstruction that is both accurate and complete.

We will use $F(\tau)$ as the default measure for the benchmark. We will also report additional measures to facilitate fine-grained analysis of the performance characteristics of each technique. In particular, we will report $P(d)$ and $R(d)$ alongside $F(d)$, across a range of distance thresholds d.

Evaluation server and leaderboard. To support progress in the field, we will set up a public evaluation server and leaderboard. Researchers will be able to submit reconstructions of benchmark sequences, which will be evaluated according to multiple measures, with $F(\tau)$ being the main measure for the leaderboard. Ground truth data for the benchmark sequences will be withheld to ensure that the reported measures reflect genuine performance characteristics of the underlying techniques. The evaluation server and leaderboard can be accessed via the benchmark's web site, www.tanksandtemples.org.

7 EVALUATED METHODS

This section describes reconstruction pipelines we have evaluated. While our benchmark is set up to evaluate complete pipelines (from video to dense point clouds), many state-of-the-art techniques focus on SfM or MVS specifically. Therefore, we have assembled many pipelines for evaluation by putting together compatible SfM and MVS methods. Sections 7.1 and 7.2 describe the specific SfM and MVS methods we have tested. Section 7.3 describes commercial software that was tested alongside open-source implementations. Section 7.4 lists all the pipelines we have evaluated. In each subsection, the methods are listed in alphabetical order.

7.1 Structure from motion

Bundler. Bundler [Snavely 2010; Snavely et al. 2008] is a seminal SfM implementation designed for large-scale community photo collections. It is the oldest SfM implementation we have tested and has inspired many related efforts in the community. To process large image collections, Bundler adds images incrementally. Starting from an image pair that has the largest number of matched feature points, Bundler triangulates matched points and adds a new image that best aligns with previously triangulated points. During the process, bundle adjustment is used to refine the camera poses and landmark positions by minimizing reprojection error. This general approach is known as incremental SfM and is followed in many other implementations.

COLMAP. COLMAP [Schönberger 2016] is a general-purpose SfM and MVS pipeline that is based on recently presented ideas [Schönberger and Frahm 2016; Schönberger et al. 2016]. COLMAP follows the incremental SfM approach used in Bundler but integrates additional verification, outlier filtering, and model selection techniques that increase the robustness of each stage in the pipeline [Schönberger and Frahm 2016].

MVE. MVE is a complete reconstruction pipeline that integrates SfM, MVS, and surface meshing [Fuhrmann et al. 2015]. The SfM implementation is structurally akin to Bundler and integrates multiple variations on each step of the pipeline.

OpenMVG. OpenMVG is a comprehensive and actively maintained open-source multi-view geometry library [Moulon et al. 2016]. We use the incremental SfM configuration for most experiments with OpenMVG, but also test the global SfM configuration (OpenMVG-G).

Theia. Theia [Sweeney 2016] is an open-source SfM library that provides recent implementations of both incremental and global SfM pipelines, informed by recent research. We have tested both configurations: global (Theia-G) and incremental (Theia-I).

VisualSFM. VisualSFM [Wu 2011] is a highly optimized incremental SfM pipeline that integrates ideas from multiple projects [Wu 2013; Wu et al. 2011].

7.2 Multi-view stereo

CMPMVS. CMPMVS is an implementation of the technique of Jancosek and Pajdla [2011]. It is one of the reference MVS techniques due to its handling of weakly textured surfaces.

COLMAP. This is an implementation of the recent work of Schönberger et al. [2016] and is part of the COLMAP library [Schönberger 2016].

MVE. This MVS implementation is based on the work of Goesele et al. [2007] and is part of the MVE library [Fuhrmann et al. 2015].

OpenMVS. OpenMVS is an actively maintained open-source MVS library that provides a set of algorithms for producing dense point clouds from localized cameras and landmarks. It can be viewed as a counterpart to OpenMVG but is interoperable with many SfM implementations.

PMVS. Patch-based multi-view stereo (PMVS) [Furukawa 2011] is a seminal MVS pipeline that was used in numerous projects and inspired subsequent MVS techniques. The implementation integrates ideas described in two papers [Furukawa et al. 2010; Furukawa and Ponce 2010]. We use the recommended combination of CMVS and PMVS2.

SMVS. Shading-aware multi-view stereo (SMVS) [Langguth et al. 2016] is a recently presented MVS technique that reasons about surface reflectance in order to increase the accuracy and completeness of dense reconstruction. In its default mode, it is not using shading-based optimization. This is the mode that was used for the evaluation.

7.3 Commercial software

We have also evaluated a commercial solution that provides a complete pipeline that admits a collection of images as input and produces a dense reconstruction as output.

Pix4D. Pix4D is a spinoff from the lab that created the influential EPFL benchmark [Strecha et al. 2008]. The company provides multiple products that support image-based reconstruction. We used the 2016 Pix4Dmapper Pro version for the evaluation.

7.4 Pipelines

We used SfM and MVS implementations described in the preceding sections to assemble 15 reconstruction pipelines for evaluation.

First, COLMAP, MVE, and Pix4D are already configured as complete pipelines, integrating the respective SfM and MVS methods. We evaluate these complete pipelines.

Beyond this, we evaluate many combinations of SfM and MVS methods that provide compatible interfaces:

- Bundler + PMVS
- MVE (SfM) + SMVS
- OpenMVG + MVE (MVS)
- OpenMVG + OpenMVS
- OpenMVG-G + OpenMVS
- OpenMVG + PMVS
- OpenMVG + SMVS
- Theia-G + OpenMVS
- Theia-I + OpenMVS
- VisualSfM + CMPMVS
- VisualSfM + OpenMVS
- VisualSfM + PMVS

For each pipeline, we used the settings recommended in the respective documentation.

8 RESULTS

Table 2 summarizes the performance of each of the 15 evaluated pipelines on each of the benchmark scenes. For each pipeline and each dataset, the table reports the F-score for the reconstruction produced by the pipeline on this dataset, using the default distance threshold τ. For each pipeline, the table also reports its mean F-score over datasets from the intermediate and advanced groups, respectively, as well as the average rank of each pipeline on each group. The average rank is defined as the average of the ranks of the pipeline over the whole group. (E.g., if a pipeline yields the highest F-score on half of the scenes and the fourth-highest on the other half, its average rank is 2.5.) The rank is a more robust summary statistic than the mean and should be regarded as the primary measure of the relative performance of a given pipeline. The mean provides an indication of the absolute performance.

COLMAP [Schönberger 2016; Schönberger and Frahm 2016; Schönberger et al. 2016] achieves the lowest rank on both the intermediate and the advanced groups. It yields the top F-score on four of the intermediate datasets and four of the advanced ones. On the intermediate group, Pix4D and OpenMVG+OpenMVS achieve similar aggregate performance to COLMAP, with Pix4D achieving the highest F-score on three of the datasets and obtaining the highest mean F-score across the group. Overall, Family appears to be the easiest intermediate dataset for existing pipelines and Horse the hardest. We attribute this to the uniform specular materials of the Horse statue and its pedestal.

On the advanced group, Pix4D achieves the second lowest rank after COLMAP, followed by OpenMVG+OpenMVS. In this group, Museum appears to be the easiest dataset. Auditorium and Palace are very challenging.

We are primarily interested not in the relative performance of the different pipelines, but in the best performance across all pipelines on each dataset. The best performance of existing techniques on a dataset indicates how much room for progress remains. Figure 5 shows the reconstructions obtained by the best-performing pipelines on a number of datasets. For each pipeline, the figure shows the color-coded reconstruction, with colors indicating per-point distance to the ground-truth model, as well as the color-coded ground-truth, with colors indicating per-point distance to the reconstruction. Additional visualizations are provided in the supplement.

Evaluating individual components. Our benchmark evaluates complete reconstruction pipelines. Individual components, such as specific SfM or MVS systems, can still be evaluated by fixing the other components. In our results, OpenMVG is used as an SfM front-end for four different MVS systems. These MVS systems can be compared to each other in this way. Likewise, five SfM systems are used with OpenMVS, which allows a comparison of these SfM systems. For example, comparing "OpenMVG + OpenMVS" to "OpenMVG-G + OpenMVS" and "Theia-I + OpenMVS" to "Theia-G + OpenMVS" reveals that incremental SfM systems outperform global SfM.

Analysis. A major factor in the performance of reconstruction pipelines is the robustness of the SfM system. SfM systems often

		Bundler + PMVS	COLMAP	MVE	MVE + SMVS	OpenMVG + MVE	OpenMVG + OpenMVS	OpenMVG-G + OpenMVS	OpenMVG + PMVS	OpenMVG + SMVS	Pix4D	Theia-G + OpenMVS	Theia-I + OpenMVS	VisualSfM + CMPMVS	VisualSfM + OpenMVS	VisualSfM + PMVS
Intermediate	Family	16.91	50.41	48.59	30.42	49.91	58.86	56.50	41.03	31.93	**64.45**	47.95	48.11	35.41	49.10	38.02
	Francis	4.34	22.25	23.84	16.64	28.19	**32.59**	29.63	17.70	19.92	31.91	19.52	19.38	14.11	21.38	12.93
	Horse	3.82	25.63	12.70	10.44	20.75	26.25	21.69	12.83	15.02	**26.43**	19.56	20.66	14.71	18.59	11.30
	Lighthouse	22.49	**56.43**	5.07	39.16	43.35	43.12	6.55	36.68	39.38	54.41	28.90	30.02	37.75	25.24	41.75
	M60	23.80	44.83	39.62	34.35	44.51	44.73	39.54	35.93	36.51	**50.58**	16.25	30.37	12.02	27.02	35.47
	Panther	21.54	**46.97**	38.16	37.90	44.76	46.85	28.48	33.20	41.61	35.37	21.54	30.79	24.29	24.64	34.19
	Playground	0.53	**48.53**	5.81	2.40	36.58	45.97	0.00	31.78	35.89	47.78	23.45	23.65	27.26	16.59	35.47
	Train	9.42	**42.04**	29.19	21.44	35.95	35.27	0.53	28.10	25.12	34.96	10.24	20.46	13.62	13.07	13.26
	Mean	12.86	42.14	25.37	24.09	38.00	41.71	22.86	29.66	30.67	**43.24**	23.43	27.93	22.40	24.45	27.80
	Rank	14.25	**2.38**	8.25	10.50	3.75	2.50	8.88	8.88	7.38	2.50	10.88	9.12	11.12	10.00	9.62
Advanced	Auditorium	0.00	**16.02**	4.11	0.97	14.70	9.79	1.89	4.54	6.96	10.83	5.74	6.23	4.70	7.94	4.68
	Ballroom	4.05	25.23	12.63	6.76	**26.36**	22.49	9.16	12.09	11.58	18.53	13.63	13.73	8.07	15.21	10.84
	Courtroom	10.30	**34.70**	27.93	16.97	32.48	26.54	24.61	21.00	19.82	33.21	16.08	18.43	13.17	21.21	16.36
	Museum	11.15	41.51	34.67	19.72	37.57	36.89	26.18	29.17	21.89	**47.37**	15.51	18.55	8.66	19.78	20.00
	Palace	2.71	**18.05**	13.58	7.74	3.65	14.64	4.02	6.76	8.90	14.47	6.43	10.61	3.89	9.10	7.32
	Temple	5.45	**27.94**	16.79	7.98	22.84	20.76	14.14	12.72	12.27	26.01	11.77	11.58	6.95	2.99	2.12
	Mean	5.61	**27.24**	18.28	10.02	22.93	21.85	13.33	14.38	13.57	25.07	11.53	13.19	7.57	12.70	10.22
	Rank	14.50	**1.33**	6.33	11.50	4.33	3.67	9.33	8.50	8.00	2.50	10.17	8.33	12.67	7.83	11.00

Table 2. F-score for each method on each benchmark dataset. Mean F-score and average rank are also listed, summarizing performance on the intermediate and advanced groups.

produce disconnected clusters that are not properly integrated to make up a complete model of the scene. Big differences in the ranking among current pipelines often come from failures on the part of the SfM system to integrate a whole scene together properly. Interestingly, all of the tested SfM systems produce camera poses for at least 80% of the input images. However, the quality of these camera poses varies drastically. In addition to producing disconnected clusters, some SfM systems produce clearly inaccurate poses that do not make up a plausible camera path.

Characteristics of the MVS system are also important. Systems that incorporate meshing remove outliers but sparsify flat regions. The performance of these methods on our benchmark can be improved by meshing at uniform density. On the other hand, systems that do not mesh or filter outliers by other means produce noisy point sets that have limited precision.

Robustness to varying exposure, specular materials, and uniform surfaces are also important. On datasets with varying exposure settings, MVS methods with more restrictive modeling assumptions are less robust. All MVS methods are challenged by uniform surfaces and strongly non-Lambertian materials.

9 CONCLUSION

We have presented a new benchmark for evaluating image-based reconstruction techniques. The presented benchmark has a number of characteristics that can support the development of new approaches to 3D reconstruction. Video sequences are provided as input, encouraging new ideas that take advantage of temporally dense sampling to increase reconstruction fidelity. Complete pipelines are evaluated, aiming to support systems that tackle camera localization and dense reconstruction jointly. Both outdoor and indoor scenes are included, with the goal of stimulating the development of robust broad-competence systems. We will set up an evaluation server and online leaderboard that can be used by the community to track progress. The datasets, evaluation server, and leaderboard can be accessed via the benchmark's web site, www.tanksandtemples.org.

ACKNOWLEDGEMENTS

We thank the anonymous reviewers for their detailed and professional reviews. Figures 1–4 were created using Mitsuba [Jakob 2010].

REFERENCES

Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. 2016. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision* 120, 2 (2016).

Sameer Agarwal, Yasutaka Furukawa, Noah Snavely, Ian Simon, Brian Curless, Steven M. Seitz, and Richard Szeliski. 2011. Building Rome in a day. *Communications of the ACM* 54, 10 (2011).

Sameer Agarwal, Noah Snavely, Steven M. Seitz, and Richard Szeliski. 2010. Bundle adjustment in the large. In *ECCV*.

Matthew Berger, Joshua A. Levine, Luis Gustavo Nonato, Gabriel Taubin, and Cláudio T. Silva. 2013. A benchmark for surface reconstruction. *ACM Transactions on Graphics* 32, 2 (2013).

Michael Burri, Janosch Nikolic, Pascal Gohl, Thomas Schneider, Joern Rehder, Sammy Omari, Markus W. Achtelik, and Roland Siegwart. 2016. The EuRoC micro aerial vehicle datasets. *International Journal of Robotics Research* 35, 10 (2016).

Sungjoon Choi, Qian-Yi Zhou, and Vladlen Koltun. 2015. Robust reconstruction of indoor scenes. In *CVPR*.

Jakob Engel, Vladlen Koltun, and Daniel Cremers. 2017. Direct sparse odometry. *Pattern Analysis and Machine Intelligence* 39 (2017).

Jakob Engel, Thomas Schöps, and Daniel Cremers. 2014. LSD-SLAM: Large-scale direct monocular SLAM. In *ECCV*.

Jan-Michael Frahm, Marc Pollefeys, Svetlana Lazebnik, David Gallup, Brian Clipp, Rahul Raguram, Changchang Wu, Christopher Zach, and Tim Johnson. 2010. Fast robust large-scale mapping from video and Internet photo collections. 65, 6 (2010).

Simon Fuhrmann, Fabian Langguth, Nils Moehrle, Michael Waechter, and Michael Goesele. 2015. MVE – An image-based reconstruction environment. *Computers & Graphics* 53 (2015).

Yasutaka Furukawa. 2011. CMVS and PMVS2. http://www.di.ens.fr/cmvs. (2011).

Yasutaka Furukawa, Brian Curless, Steven M. Seitz, and Richard Szeliski. 2009. Reconstructing building interiors from images. In *ICCV*.

Yasutaka Furukawa, Brian Curless, Steven M. Seitz, and Richard Szeliski. 2010. Towards Internet-scale multi-view stereo. In *CVPR*.

Yasutaka Furukawa and Carlos Hernández. 2015. Multi-view stereo: A tutorial. *Foundations and Trends in Computer Graphics and Vision* 9, 1-2 (2015).

Yasutaka Furukawa and Jean Ponce. 2010. Accurate, dense, and robust multiview stereopsis. *Pattern Analysis and Machine Intelligence* 32, 8 (2010).

Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. 2013. Vision meets robotics: The KITTI dataset. *International Journal of Robotics Research* 32, 11 (2013).

Michael Goesele, Noah Snavely, Brian Curless, Hugues Hoppe, and Steven M. Seitz. 2007. Multi-view stereo for community photo collections. In *ICCV*.

Ankur Handa, Thomas Whelan, John McDonald, and Andrew J. Davison. 2014. A benchmark for RGB-D visual odometry, 3D reconstruction and SLAM. In *ICRA*.

Richard Hartley and Andrew Zisserman. 2000. *Multiple view geometry in computer vision*. Cambridge University Press.

Jared Heinly, Johannes L. Schönberger, Enrique Dunn, and Jan-Michael Frahm. 2015. Reconstructing the world* in six days. In *CVPR*.

Satoshi Ikehata, Hang Yang, and Yasutaka Furukawa. 2015. Structured indoor modeling. In *ICCV*.

Wenzel Jakob. 2010. Mitsuba renderer. http://www.mitsuba-renderer.org. (2010).

Michal Jancosek and Tomas Pajdla. 2011. Multi-view reconstruction preserving weakly-supported surfaces. In *CVPR*.

Kalin Kolev, Petri Tanskanen, Pablo Speciale, and Marc Pollefeys. 2014. Turning mobile phones into 3D scanners. In *CVPR*.

Fabian Langguth, Kalyan Sunkavalli, Sunil Hadap, and Michael Goesele. 2016. Shading-aware multi-view stereo. In *ECCV*.

Xiaowei Li, Changchang Wu, Christopher Zach, Svetlana Lazebnik, and Jan-Michael Frahm. 2008. Modeling and recognition of landmark image collections using iconic scene graphs. In *ECCV*.

Andrew Mastin, Jeremy Kepner, and John Fisher. 2009. Automatic registration of LIDAR and optical images of urban scenes. In *CVPR*.

Paul Merrell, Philippos Mordohai, Jan-Michael Frahm, and Marc Pollefeys. 2007. Evaluation of large scale scene reconstruction. In *ICCV Workshops*.

Pierre Moulon, Pascal Monasse, Renaud Marlet, and others. 2016. OpenMVG: An open multiple view geometry library. https://github.com/openMVG/openMVG. (2016).

Richard A. Newcombe, Steven Lovegrove, and Andrew J. Davison. 2011. DTAM: Dense tracking and mapping in real-time. In *ICCV*.

Marc Pollefeys, David Nistér, Jan-Michael Frahm, Amir Akbarzadeh, Philippos Mordohai, Brian Clipp, Chris Engels, David Gallup, Seon Joo Kim, Paul Merrell, C. Salmi, Sudipta N. Sinha, B. Talton, Liang Wang, Qingxiong Yang, Henrik Stewénius, Ruigang Yang, Greg Welch, and Herman Towles. 2008. Detailed real-time urban 3D reconstruction from video. *International Journal of Computer Vision* 78, 2-3 (2008).

Johannes L. Schönberger. 2016. COLMAP. https://colmap.github.io. (2016).

Johannes L. Schönberger and Jan-Michael Frahm. 2016. Structure-from-motion revisited. In *CVPR*.

Johannes L. Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. 2016. Pixelwise view selection for unstructured multi-view stereo. In *ECCV*.

Thomas Schöps, Torsten Sattler, Christian Häne, and Marc Pollefeys. 2015. 3D modeling on the go: Interactive 3D reconstruction of large-scale scenes on mobile devices. In *3DV*.

Thomas Schöps, Johannes L. Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. 2017. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *CVPR*.

Steven M. Seitz, Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski. 2006. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *CVPR*.

Qi Shan, Riley Adams, Brian Curless, Yasutaka Furukawa, and Steven M. Seitz. 2013. The visual Turing test for scene reconstruction. In *3DV*.

Noah Snavely. 2010. Bundler: Structure from motion (SfM) for unordered image collections. https://github.com/snavely/bundler_sfm. (2010).

Noah Snavely, Steven M. Seitz, and Richard Szeliski. 2008. Modeling the world from Internet photo collections. *International Journal of Computer Vision* 80, 2 (2008).

Christoph Strecha, Wolfgang von Hansen, Luc J. Van Gool, Pascal Fua, and Ulrich Thoennessen. 2008. On benchmarking camera calibration and multi-view stereo for high resolution imagery. In *CVPR*.

Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. 2012. A benchmark for the evaluation of RGB-D SLAM systems. In *IROS*.

Chris Sweeney. 2016. Theia multiview geometry library. http://theia-sfm.org. (2016).

Petri Tanskanen, Kalin Kolev, Lorenz Meier, Federico Camposeco, Olivier Saurer, and Marc Pollefeys. 2013. Live metric 3D reconstruction on mobile phones. In *ICCV*.

Engin Tola, Christoph Strecha, and Pascal Fua. 2012. Efficient large-scale multi-view stereo for ultra high-resolution image sets. *Machine Vision and Applications* 23, 5 (2012).

Bill Triggs, Philip Mclauchlan, Richard Hartley, and Andrew Fitzgibbon. 2000. Bundle adjustment – a modern synthesis. In *Vision Algorithms: Theory and Practice*.

Shinji Umeyama. 1991. Least-squares estimation of transformation parameters between two point patterns. *Pattern Analysis and Machine Intelligence* 13, 4 (1991).

George Vogiatzis and Carlos Hernández. 2011. Video-based, real-time multi-view stereo. *Image and Vision Computing* 29, 7 (2011).

Hoang-Hiep Vu, Patrick Labatut, Jean-Philippe Pons, and Renaud Keriven. 2012. High accuracy and visibility-consistent dense multiview stereo. *Pattern Analysis and Machine Intelligence* 34, 5 (2012).

Michael Waechter, Mate Beljan, Simon Fuhrmann, Nils Moehrle, Johannes Kopf, and Michael Goesele. 2017. Virtual rephotography: Novel view prediction error for 3D reconstruction. *ACM Transactions on Graphics* 36, 1 (2017).

Andreas Wendel, Michael Maurer, Gottfried Graber, Thomas Pock, and Horst Bischof. 2012. Dense reconstruction on-the-fly. In *CVPR*.

Changchang Wu. 2011. VisualSFM: A visual structure from motion system. http://ccwu.me/vsfm. (2011).

Changchang Wu. 2013. Towards linear-time incremental structure from motion. In *3DV*.

Changchang Wu, Sameer Agarwal, Brian Curless, and Steven M. Seitz. 2011. Multicore bundle adjustment. In *CVPR*.

Jianxiong Xiao and Yasutaka Furukawa. 2014. Reconstructing the world's museums. *International Journal of Computer Vision* 110, 3 (2014).

Qian-Yi Zhou and Vladlen Koltun. 2013. Dense scene reconstruction with points of interest. *ACM Transactions on Graphics* 32, 4 (2013).

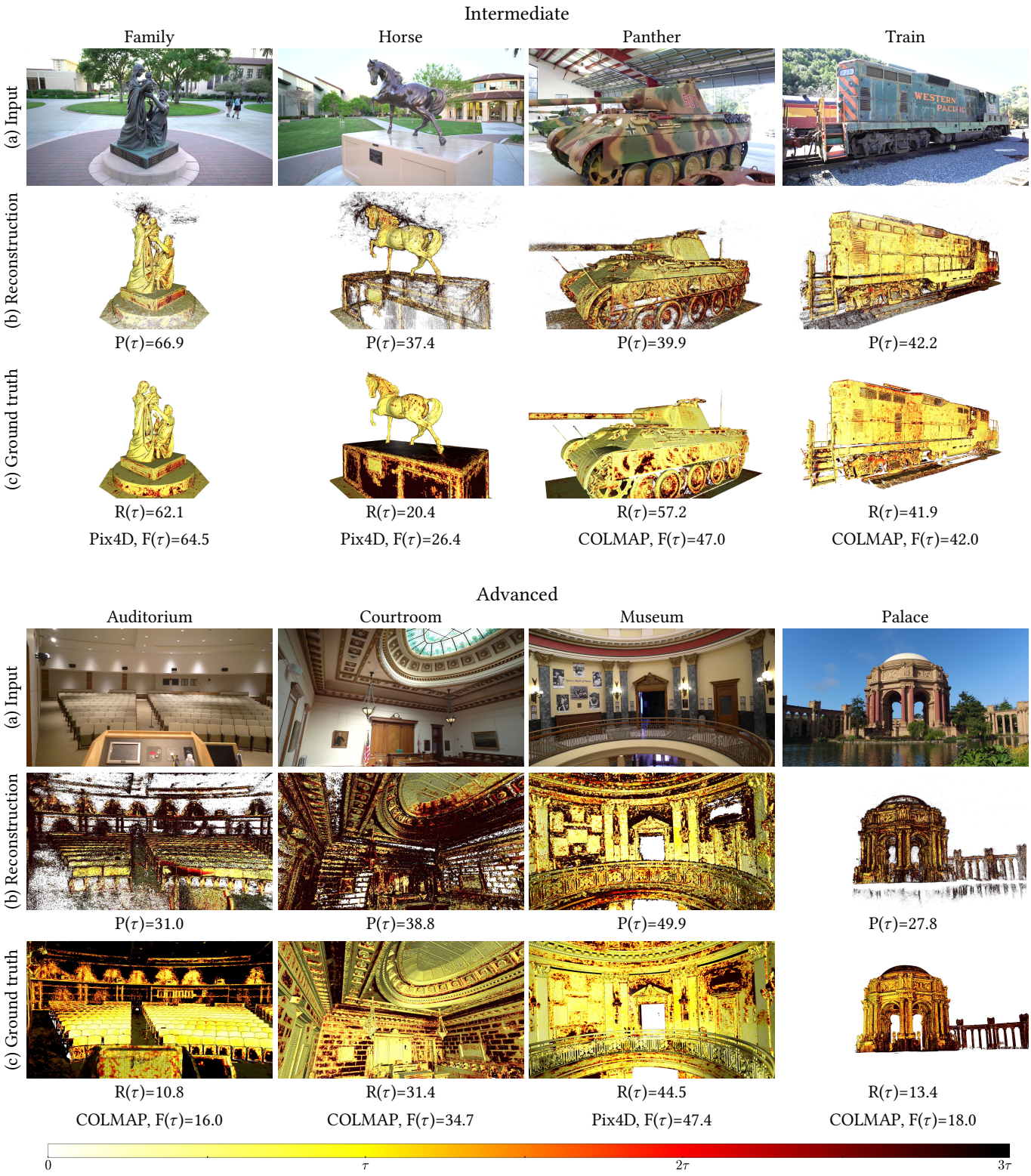

Fig. 5. State-of-the-art results on a number of benchmark datasets. (a) A frame from the input video sequence. (b) Reconstruction produced by the best-performing pipeline, with distance to the ground-truth model coded by color. (c) The ground-truth model, with per-point distance to the reconstruction coded by color.