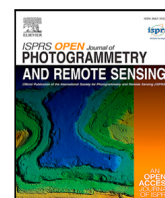




Contents lists available at ScienceDirect

ISPRS Open Journal of Photogrammetry and Remote Sensing

journal homepage: www.journals.elsevier.com/isprs-open-journal-of-photogrammetry-and-remote-sensing

JB3D: An aerial video dataset with per-frame camera poses and LiDAR ground truth for multi-view stereo and novel view synthesis benchmarking[☆]

M. Hermann^{a,b}, D. Böhmer^a, S. Hinz^b, M. Weinmann^b^a Fraunhofer IOSB, Karlsruhe, Germany^b Institute of Photogrammetry and Remote Sensing, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

ARTICLE INFO

Keywords:

Dataset
Aerial
Video
LiDAR
3D reconstruction
MVS
NVS

ABSTRACT

Image-based 3D reconstruction from unmanned aerial vehicles is crucial for many applications like navigation, environmental monitoring, or creating digital twins. However, real-world aerial footage introduces unconstrained motion, scale variation and dynamic elements that challenge current methods. Existing benchmarks address many of these aspects while providing valuable LiDAR ground truth, but only a few ground-based benchmarks provide video data with camera poses, often operating at lower frame rates or lower image resolutions, and most aerial benchmarks primarily offer still images only. To address this issue, we introduce JB3D: an aerial video dataset designed for benchmarking multi-view stereo and novel view synthesis approaches in realistic outdoor environments. Captured over the course of a two-day outdoor festival featuring challenging scenery, including fine structures and dynamic elements such as moving cars and people, JB3D comprises 14 UAV videos at 30 FPS, five LiDAR scans, high-resolution RGB and thermal infrared images. To accurately reflect real-world operating conditions such as infrastructure monitoring, crowd counting, or supporting emergency response forces, all flight trajectories for video acquisition were controlled manually. All the data are registered in a single ECEF/ENU coordinate frame. The novelty compared to other benchmark datasets relates to the per-frame camera poses, for which reference data are provided, too. They enable reconstruction from individual sequences, multi-sequence evaluation and cross-day change detection. Alongside baseline results from conventional, learning-based, and advanced scene representation approaches, we release evaluation and visualization routines (<https://github.com/JB3D-Dataset/Multi-View-Stereo>). In addition, we provide insights into how the strengths and weaknesses of chosen evaluation metrics affect the outcome by employing established metrics, such as L1-abs, F1-Score and Chamfer distance.

1. Introduction

The field of image-based 3D reconstruction has been advancing rapidly, particularly with the advent of deep learning techniques. A crucial role can be attributed to the processing 2D image and video data into detailed 3D reconstructions, which are essential for a variety of applications, including autonomous navigation, urban and environmental monitoring, as well as creating digital twins. Commercial off-the-shelf unmanned aerial vehicles (UAVs) are vital in this context, as they provide fast and high-precision aerial imaging capabilities at various scales, enabling efficient data acquisition for diverse applications such as disaster response, agriculture, and urban planning. However, the acquisition of reliable three-dimensional data from aerial imagery

remains challenging due to factors such as unconstrained flight trajectories, diverse scales and motions, and the dynamic nature of outdoor environments. These factors frequently violate assumptions of standard MVS pipelines, but recent innovations, including learning-based MVS approaches and approaches based on advanced scene representations, have improved in this regard.

Existing datasets address various aspects for evaluating the quality of image-based 3D reconstruction, yet do not sufficiently address many critical requirements for comprehensive performance evaluation under real-world conditions for UAVs. For instance, the Middlebury (Scharstein and Szeliski, 2002) and DTU (Jensen et al., 2014) datasets are ideal for benchmark studies involving small-scale objects with controlled lighting, whereas ETH3D (Schöps et al., 2017) and Tanks

[☆] Given his role as EIC, Dr. Martin Weinmann had no involvement in the peer review of this article and had no access to information regarding its peer review. Full responsibility for the editorial process for this article was delegated to another journal editor.

* Corresponding author at: Fraunhofer IOSB, Karlsruhe, Germany.

E-mail addresses: max.hermann@iosb.fraunhofer.de, max.hermann@kit.edu (M. Hermann), dennis.boehmer@iosb.fraunhofer.de (D. Böhmer), stefan.hinz@kit.edu (S. Hinz), martin.weinmann@kit.edu (M. Weinmann).

<https://doi.org/10.1016/j.ophoto.2026.100149>

Received 12 January 2026; Received in revised form 7 August 2026; Accepted 12 August 2026

Available online 6 September 2026

2667-3932/© 2026 The Authors. Published by Elsevier B.V. on behalf of International Society of Photogrammetry and Remote Sensing. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>). Open access funding enabled and organized by Projekt DEAL.

and Temples (Knapitsch et al., 2017) provide high-resolution and large-scale outdoor scenes with a focus on more realistic conditions. Tanks and Temples is particularly noteworthy because it provides video data with 1 FPS ground truth camera trajectories. However, despite this, these datasets only provide terrestrial scenes without any additional metadata. Synthetic datasets, such as BlendedMVS (Yao et al., 2020), have been shown to effectively expand the scale for training complex models, but they have been observed to struggle to replicate the conditions and complexities inherent to real-world aerial data. Whilst the UseGeo dataset (Nex et al., 2024; Hermann et al., 2024b) has made progress in providing a UAV-specific benchmarks with LiDAR-based ground truth, its focus is on static still imagery.

To overcome the above-mentioned shortcomings, we introduce JB3D, a georeferenced aerial video dataset with per-frame camera poses and LiDAR ground truth designed to evaluate the performance of image-based 3D reconstruction approaches. To accurately reflect real-world operating conditions such as infrastructure monitoring, crowd counting, or supporting emergency response forces, all flight trajectories for video acquisition were controlled manually. The dataset captures a two-day outdoor festival setup with changing geometry, dynamic elements such as people and vehicles and fine structures like fences and vegetation. It comprises 14 UAV videos at 30FPS leading to approx. 50,000 video frames, 5 LiDAR scans, approx. 5000 high-resolution RGB still images, and approx. 3000 thermal still images. All data is registered in one common ECEF/ENU coordinate frame, enabling 3D reconstruction of single video sequences, evaluation using multiple sequences or cross-day change detection.

Our contributions can be summarized as follows:

- We release a comprehensive dataset consisting of 14 UAV-based videos with metadata and estimated frame-accurate camera poses, as well as corresponding LiDAR point clouds for evaluating the performance of image-based 3D reconstruction approaches. The entire dataset uses a single joined ECEF/ENU coordinate frame.
- In addition, the dataset provides 5 videos that are less suitable for evaluating the performance of image-based 3D reconstruction approaches and 5000 RGB and 3000 thermal still images from nadir and oblique perspective, resulting in a total of approximately 55,000 RGB images. The intrinsic and extrinsic camera calibrations are provided for all RGB images.
- We provide an extensive comparison of different baseline approaches, consisting of conventional and deep learning approaches, as well as results achieved by approaches relying on advanced scene representations. In order to replicate our results, we are also releasing the evaluation and visualization scripts used for this purpose ¹.

This paper is structured as follows. In Section 2, we summarize the current state-of-the-art regarding existing datasets, categorizing them according to their characteristics, and provide an overview of conventional and more recent approaches for image-based 3D reconstruction methods. Next, in Section 3, we present our approach to conducting the measurement campaign and how we performed the post-processing to create the dataset. In this section we also introduce the evaluation metrics, which are provided as a script¹. Section 4 presents the quantitative and qualitative results of four baseline approaches in different categories. These results are then discussed in Section 5, alongside the impact of the choice of evaluation metric. We finish by summarizing the results in Section 6 and providing concluding remarks.

2. Related work

In the following, we briefly summarize related work in terms of existing datasets, and approaches for image-based 3D reconstruction. For the latter, we focus on traditional approaches, standard learning-based approaches, and learning-based approaches relying on advanced scene representations. Table 1 shows an overview of existing image-based datasets intended for 3D reconstruction benchmarking.

2.1. Datasets

Early work on evaluating the performance of MVS reconstruction algorithms resulted in the Middlebury MVS benchmark (Seitz et al., 2006) addressing a comparison of derived MVS reconstructions with highly accurate ground truth 3D models acquired via laser scanning. For each of two considered objects, images were acquired with three different view constellations (i.e., either views sampled on a hemisphere or views sampled on a ring around the object), thus yielding six different image sets as the basis for 3D reconstruction. Since the considered objects are rather small and acquired images of low resolution, a benchmark for image-based 3D reconstruction with a focus on outdoor scenes and high-resolution imagery was proposed (Strecha et al., 2008), where the reference was acquired with a laser scanner. To account for more diversity in terms of scene characteristics (i.e., beyond well-textured diffuse-reflectance surfaces), the DTU dataset (Jensen et al., 2014; Aanaes et al., 2016) was presented. This dataset was acquired using a structured light scanner mounted on an industrial robot arm. The released images were taken by one of the cameras in the structured light scanner, while the scanner allows for capturing a reference 3D surface geometry of the considered scene/object. In total, the dataset comprises 80 small-scale scenes of large variability in terms of varying reflectance behavior, texture, and geometric properties as well as different lighting conditions.

To address more realistic scenarios in the form of challenging large-scale outdoor and indoor environments, the Tanks and Temples dataset (Knapitsch et al., 2017) has been acquired, which comprises high-resolution video sequences as well as corresponding ground truth point clouds collected with a laser scanner. This dataset has been proposed for the joint evaluation of both structure-from-motion (SfM) and MVS, i.e., the complete 3D reconstruction pipeline from the input video to the dense 3D point cloud as output. A complementary dataset has been presented with the ETH3D dataset (Schöps et al., 2017). It also covers a variety of challenging indoor and outdoor scenes for which high-resolution imagery have been acquired with a professional DSLR camera and ground truth 3D point clouds have been captured with a laser scanner. However, it additionally contains synchronized low-resolution stereo videos with varying fields-of-view. Such a consideration is more realistic for mobile devices as used in robotics, and respective videos enable algorithms to exploit redundancy across imagery acquired with high frame rates to improve the reconstruction quality. Instead of relying on expensive laser scanners for acquiring ground truth 3D point clouds, the BlendedMVS dataset (Yao et al., 2020) is a large-scale synthetic dataset which relies on the generation of training images and corresponding depth maps by rendering textured 3D models to different view points. This dataset comprises 113 textured 3D models with a variety of different scenes, including cities, architectures, sculptures and small objects, that are represented with 20 to 1000 input images, respectively. In total, more than 17,000 images are provided.

A particular scenario is given for the reconstruction of real-world outdoor urban scenes, when aerial imagery is involved. In this regard, the ISPRS benchmark for multi-platform photogrammetry (Nex et al., 2015) comprises datasets for two urban test areas in the city of Dortmund, Germany. These datasets contain nadir and oblique aerial imagery, and convergent and redundant terrestrial imagery together

¹ <https://github.com/JB3D-Dataset/Multi-View-Stereo>

Table 1

Comparison of MVS datasets commonly used for 3D reconstruction. The JB3D aerial dataset is unique in that it provides both LiDAR ground truth and videos along 30 FPS camera trajectories.

Dataset	Environment	Perspective	Modality	Ground truth	Camera poses
Middlebury MVS (Seitz et al., 2006)	Lab	Ground	Images	LiDAR	yes
DTU (Jensen et al., 2014)	Lab	Aerial/Ground	Images	Structured-light	yes
Benchmark for multi-platform photogrammetry (Nex et al., 2015)	Outdoor	Aerial/Ground	Images	LiDAR	yes
BlendedMVS (Yao et al., 2020)	Synthetic	Aerial/Ground	Images	Photogrammetric	yes
UseGeo (Nex et al., 2024; Hermann et al., 2024b)	Outdoor	Aerial	Images	LiDAR	yes
ETH3D MVS (Schöps et al., 2017)	Indoor/Outdoor	Ground	Images/Low resolution videos	LiDAR	yes (13.6 FPS)
Tanks & Temples (Knapitsch et al., 2017)	Indoor/Outdoor	Ground	Images/Videos	LiDAR	yes (1 FPS)
JB3D (Ours)	Outdoor	Aerial	Images (RGB & thermal infrared)/Videos	LiDAR	yes (30 FPS)

with ground truth data acquired via airborne and terrestrial laser scanning or represented by single Ground Control Points (GCPs) as well as single target coordinates. The extension from rather small-scale scenes to large-scale urban areas has been addressed with the UseGeo dataset (Nex et al., 2024; Hermann et al., 2024b) which has been released as a large-scale real-world dataset allowing for quantitative and qualitative evaluation of approaches for image-based depth estimation and 3D reconstruction from UAV-borne imagery. The dataset covers an area of $1100 \text{ m} \times 650 \text{ m}$, for which aerial imagery (GSD of approximately 2 cm) and LiDAR data (point density of about 51.39 points/m) have been acquired from an average flight altitude of 80 m above ground. The acquired LiDAR data can serve as a reference, while the imagery can be used as input for depth estimation and 3D reconstruction approaches. The UseGeo dataset has been integrated in the NeRFBK dataset (Yan et al., 2023) representing a repository of real and synthetic scenes with industrial objects, transparent and reflective objects, aerial scenes and heritage scenarios. For each case, both imagery and accurate ground truth 3D models are provided.

2.2. Traditional approaches

Traditional approaches for image-based 3D reconstruction rely on the use of hand-crafted features and matching metrics to achieve multi-view consistency. PhotoTourism/Bundler (Snavely et al., 2006) represents one of the first open-source SfM pipelines, and it could even be applied on imagery without prior knowledge about intrinsic camera parameters. Since only local features between images are matched and triangulated, the achieved 3D reconstruction is represented by a rather sparse point cloud. Hence, the results of an SfM approach are typically used as input to a subsequent MVS approach such as PMVS/CMVS (Furukawa and Ponce, 2009) and numerous follow-up developments (Stathopoulou and Remondino, 2023) in order to achieve a dense 3D reconstruction. Meanwhile, numerous open-source image-based 3D reconstruction pipelines (Stathopoulou et al., 2019; Ruano and Smolic, 2021) cover both SfM and MVS, such as COLMAP (Schönberger and Frahm, 2016; Schönberger et al., 2016), which is actively maintained and easy-to-use for both expert and non-expert users.

Traditional approaches heavily rely on the assumption of photometric consistency, and the reconstruction of weakly-textured or reflective surfaces remains problematic (Huang et al., 2018). To address this, a multi-scale geometric consistency guided and planar prior assisted MVS approach has been presented with ACMMP (Xu et al., 2022).

2.3. Standard learning-based approaches

Learning-based approaches for image-based 3D reconstruction can be categorized with respect to the given scene representation (Wang et al., 2021; Stathopoulou and Remondino, 2023). Voxel-based approaches focus on the direct prediction of globally coherent voxel-occupancy grids from the given input imagery. Such approaches are constrained by the predefined resolution of a reconstructed object or scene and thus not suitable for use cases focusing on large-scale 3D reconstruction. Depth-map-based approaches focus on the prediction of

a depth map for each view and a subsequent fusion of all predicted depth maps into a coherent 3D representation. Due to this decoupling, such approaches are more efficient and thus also suitable for use cases focusing on large-scale 3D reconstruction.

In this paper, we focus on large-scale 3D reconstruction and hence depth-map-based approaches are more promising. Respective approaches typically rely on deep networks, e.g., DeepMVS (Huang et al., 2018) and MVSNet (Yao et al., 2018) and numerous variants thereof, which perform optimization based on 3D cost volumes. For instance, UniMVSNet (Peng et al., 2022) is a coarse-to-fine framework which directly constrains the cost volume like classification methods, but also realizes sub-pixel depth prediction like regression methods. GeoMVSNet (Zhang et al., 2023) integrates coarse geometric structures into finer depth estimations for improved geometry awareness. The most recent learning-based approaches for image-based 3D reconstruction also comprise transformer-based approaches such as TransMVSNet (Ding et al., 2022), MVSFormer (Cao et al., 2022) or MVSFormer++ (Cao et al., 2024).

2.4. Learning-based approaches relying on advanced scene representations

Recently, a lot of attention has been paid to implicit neural scene representations, where the underlying idea consists in representing a scene in terms of the weights of a neural network. These weights are optimized in a supervised manner to match the predicted scene appearance under certain views to the respectively observed input image for the corresponding view configurations. Consequently, it can be assumed that scene characteristics are adequately represented in the neural scene model if the deviations to the input reference views are small.

In particular, Neural Radiance Fields (NeRFs) (Mildenhall et al., 2021) and their many extensions (Tewari et al., 2022) offer a high potential for accurate scene representation. Originally proposed for the task of novel view synthesis, such approaches rely on a neural network in order to predict both density and view-dependent color information for each point in the volume. On this basis, volume rendering techniques can be used to synthesize new images from the volumetric scene information. Accurate geometric models can be achieved by NeRF extensions towards improving rendering quality (and thus also model quality) by reducing aliasing (Barron et al., 2021, 2022, 2023) and by increasing robustness to inconsistencies (Sabour et al., 2023; Buschmann et al., 2024), whereas improved efficiency can be achieved by accelerating the training of the underlying network (Müller et al., 2022; Chen et al., 2022). Furthermore, NeRF extensions have been presented that are dedicated to handle unconstrained image collections (Martin-Brualla et al., 2021; Lin et al., 2021), sparse views as input (Yu et al., 2021; Niemeyer et al., 2022; Chen et al., 2024) or large-scale scenarios (Tancik et al., 2022; Turki et al., 2022; Xiangli et al., 2022; Mi and Xu, 2023; Xie et al., 2023; Xu et al., 2024; Hermann et al., 2024a).

While NeRFs hold great potential for numerous applications, the involved neural network lacks interpretability and the extraction of object surfaces from the implicit surface information requires network

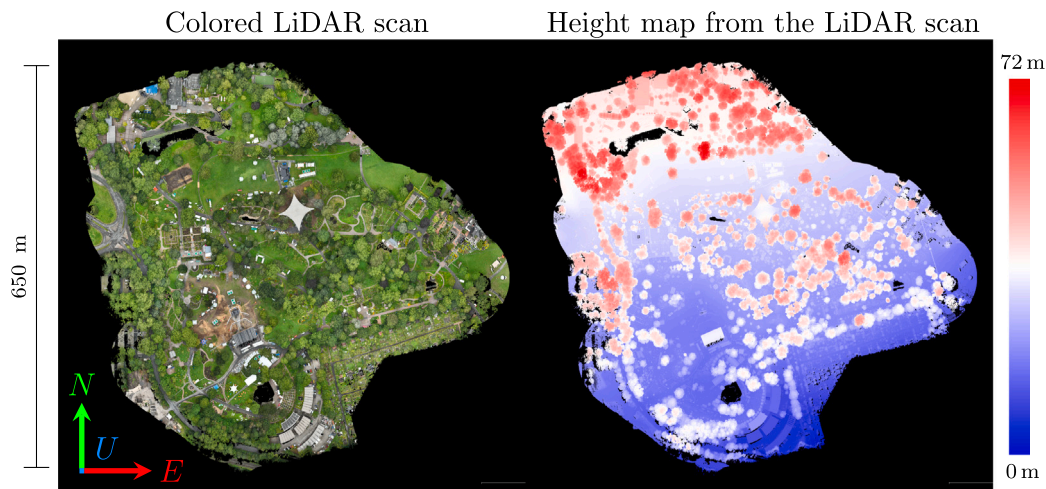


Fig. 1. Overview of the considered area. The left side shows one colored LiDAR scan. On the right side, the color-coded height map is displayed.

evaluations on a dense grid. For this reason, many recent approaches rely on an explicit surface representation. In particular, 3D Gaussian Splatting (3DGS) (Kerbl et al., 2023) has become the state-of-the-art learning-based approach for image-based 3D reconstruction. 3DGS relies on the use of anisotropic 3D Gaussians as an explicit scene representation and their optimization for novel view synthesis. Among a diversity of extensions of the original 3DGS, the fast planar-based Gaussian splatting reconstruction representation (PGSR) (Chen et al., 2025) has been proposed to achieve high-fidelity surface reconstruction while ensuring high-quality image rendering.

3. Methodology and dataset characteristics

In the following section, we discuss the characteristics of the presented dataset in detail and explain our procedure for data acquisition and post-processing. For this purpose, we first provide a quantitative and qualitative overview of the individual scenes, before describing the evaluated approaches for image-based 3D reconstruction and the metrics used to analyze the resulting point clouds. Fig. 1 shows an overview of the entire area: on the left is one colored LiDAR point cloud, and on the right is the color-coded height map generated from the LiDAR scan displayed. The entire area measures approximately 650 by 750 m.

3.1. Overview of the dataset

The dataset consists of three types of acquisitions that serve complementary purposes: (i) LiDAR scans providing the ground truth geometry, (ii) still RGB and thermal infrared images that support the data processing and offer an additional modality, and (iii) manually flown high-resolution videos, which are the primary data used for all subsequent analysis. All acquisitions are registered in a single common ECEF/ENU coordinate frame.

The LiDAR scans serve as the reference geometry for all evaluations. They were captured with a DJI Matrice carrying the DJI L2 sensor and processed with DJI Terra. LiDAR data was acquired in five separate flights over the two days, covering the whole area on each day as well as the higher-density subareas around the main stage and the second stage. Table 3 provides an overview of the resulting five point clouds, including their coverage, acquisition time, number of points, and point density. It should be noted that the scan *Whole area 1* contains fewer points, because problems with the RTK metadata led to issues during its post-processing.

In parallel to the LiDAR acquisition, a DJI Mavic3t captured still images, recording RGB and thermal infrared images simultaneously. These

images were acquired using an automatic double-grid flight pattern at a flight altitude of up to 100 m, from both nadir and oblique perspectives. The RGB images have a resolution of up to 5280×3956 pixels, whereas the thermal infrared images have a lower resolution of 640×512 pixels and a different field of view, visible in Fig. 5. The still images serve two supporting roles: they are used by DJI Terra to colorize the LiDAR point clouds, and they are included in the sparse reconstructions to improve and stabilize the video-based trajectory estimation. In total, approximately 5000 RGB and 3000 thermal still images are provided. These still images are not evaluated as reconstruction inputs themselves. They are released for completeness and as an additional modality for future work.

The primary focus of the dataset is on the manually flown videos, and all subsequent analysis in this paper is performed solely on these respective RGB videos. All videos were recorded with the DJI Mavic3t at an altitude between 60 and 80 m and have a resolution of 3840×2160 pixels at 30 FPS. To accurately reflect real-world operating conditions such as infrastructure monitoring, crowd counting, or supporting emergency response forces, all video trajectories were controlled manually. Table 2 provides an overview of the 14 videos, which differ considerably in terms of translation and rotation speed. Fig. 2 shows the corresponding 30 FPS camera trajectories projected onto the XY plane, together with the relevant section of the ground truth LiDAR scan. As can be seen, most videos are orbital flights around one or more objects of interest in the center. Since all flights were performed manually, they do not follow a perfect circular path and vary in speed and camera angle. All parts of the scene were covered at least once on both days, but with different flight paths. For evaluation, we crop the point clouds using manually defined polygons that are based on the covered area, the flight altitude, and the camera angle.

All global navigation satellite system (GNSS) metadata was recorded onboard with the help of a RTK ground station. In addition to these raw acquisitions, the dataset provides georeferenced camera poses for each still image and for all 30 FPS video frames. Together with the common coordinate frame, this enables reconstruction from individual video sequences, multi-sequence evaluation, and cross-day change detection.

3.2. Scene characteristics

Since the dataset represents real-world scenarios, there are many challenging elements visible. Fig. 3 shows examples that include fine structures such as metal fences, communication masts and flag poles, as well as vegetation and occasional water surfaces. Because people are actively working during the data acquisition, not only does the scene change slowly over the two days, but dynamic elements such as

Table 2

Overview of the characteristics of the 14 provided videos to evaluate 3D reconstruction. The detailed flight patterns are shown in Fig. 2. The 5 videos in the bottom section are not used for 3D evaluation because they either have very challenging camera movement or are located outside of the LiDAR ground truth.

Scene	Day	Time	No. images	Velocity in m/s	Angular velocity in degree/s
Main stage I	Day 1	17:31	1947	3.46 ± 1.79	3.05 ± 1.59
Main stage II	Day 1	17:37	2539	9.50 ± 2.46	8.76 ± 2.61
Flight to entrance	Day 1	17:39	1015	9.10 ± 4.72	3.19 ± 5.63
Entrance I	Day 1	17:44	5591	4.51 ± 2.62	3.26 ± 3.22
Second stage I	Day 1	18:02	2194	6.07 ± 2.01	5.27 ± 2.87
Garden I	Day 1	18:04	3554	4.92 ± 1.32	4.01 ± 1.43
Food stalls I	Day 1	18:07	2533	4.93 ± 1.16	4.46 ± 1.61
Garden II	Day 1	18:12	1314	8.38 ± 2.51	9.35 ± 3.35
Main stage III	Day 2	12:40	3125	3.56 ± 1.92	3.23 ± 2.11
Entrance II	Day 2	12:49	2918	6.03 ± 2.45	5.46 ± 2.22
Entrance III	Day 2	12:51	2767	4.27 ± 2.06	4.79 ± 2.84
Second stage II	Day 2	12:54	4444	3.68 ± 1.93	2.59 ± 1.91
Food stalls II	Day 2	12:57	1396	10.16 ± 2.03	7.38 ± 2.24
Gardens III	Day 2	12:59	1415	8.46 ± 2.82	8.89 ± 3.76
Main stage area I	Day 1	17:33	1182	4.35 ± 5.42	9.59 ± 17.68
Main stage area II	Day 1	17:35	1026	3.72 ± 3.64	4.66 ± 4.19
Flight to tower	Day 1	17:47	834	6.95 ± 4.40	15.72 ± 21.01
Tower	Day 1	17:48	2575	4.87 ± 2.19	2.80 ± 1.86
Park area	Day 1	17:54	4070	6.48 ± 3.19	7.35 ± 9.94

Table 3

LiDAR point clouds captured with the DJI L2 sensor. The scene *Whole area I* has fewer points because we had problems with the RTK metadata, which led to issues during post-processing.

Coverage	Day	Time	No. points	Area in m ²	Density in pts/m ²
Whole area	Day 1	15:30	251,449,993	337,201	745.70
Second stage	Day 1	16:40	326,096,948	132,086	2468.83
Whole area	Day 2	11:00	370,839,108	330,526	1121.96
Main stage	Day 2	11:28	562,934,693	220,658	2551.16
Second stage	Day 2	12:09	245,885,183	126,484	1943.99

moving vehicles or people exist. Fig. 4 shows one particularly example of dynamic elements. Throughout our experiments, we did not mask any dynamic elements for the 3D reconstruction or evaluation routine. In addition to the videos, still images were also acquired. For the majority of these, thermal infrared images were also captured at the same time using the DJI Mavic3t. As can be seen in Fig. 5, however, the FOV differs slightly, and thermal infrared images have lower resolution of 640×512 pixels.

3.3. Data processing

As all the data was collected at different times but in the same geographical area, we can establish a common coordinate system for all the processed sensor data. The data is referenced in the ECEF/ENU coordinate system, which consists of a translation vector describing the position in the global ECEF system and a local coordinate system in ENU convention. This is aligned tangentially to the Earth's surface, with the X-axis pointing east, the Y-axis pointing north, and the Z-axis pointing up. This decomposition enables direct processing with conventional reconstruction and visualization tools without losing the geographical context. The densest LiDAR scan is selected as the origin for the local coordinate system, and other LiDAR scans, videos, and still images are registered relative to this scan and are therefore located in the same ECEF/ENU frame.

In this way, flights at different times or recorded on different days over the same location can be directly compared with each other. To achieve this, we use a semi-automatic process for georeferencing both the video frames and the still images. The goal is to obtain an extrinsic camera pose for every image expressed in the common ECEF/ENU frame defined by the reference LiDAR scan. For the video frames, our pipeline consists of the following steps for each individual scene:

Time synchronization We use IVS3D (Hermann et al., 2021)² to synchronize all videos with their metadata and export individual frames with the corresponding GNSS positions written to the EXIF metadata.

Sparse reconstruction Every tenth frame is used to compute a sparse SfM reconstruction using COLMAP (Schönberger and Frahm, 2016), jointly estimating the intrinsic and extrinsic camera parameters. We use up to 20,000 extracted and 40,000 matched features at the original image resolution. To support the reconstruction, additional nadir and oblique still images from the double-grid flights are added, selected based on their GNSS position. Point correspondences with a reprojection error greater than 0.3 pixels are discarded.

Similarity transformation The SfM reconstruction is initially defined in an arbitrary local coordinate frame. Using the per-image GNSS positions extracted from the EXIF metadata, we estimate a similarity transformation with 7 degrees of freedom that maps the reconstruction into the joined ENU frame. This step provides the initial georeferencing of all cameras.

Fine alignment to the LiDAR reference The resulting sparse point cloud still exhibits a small rigid offset with respect to the LiDAR reference. We therefore refine the alignment with a rigid Iterative-Closest-Point (ICP) registration (Besl and McKay, 1992) between the sparse SfM cloud and the reference LiDAR scan.

Dense pose registration Finally, all of the remaining video frames are registered using COLMAP against the sparse reconstruction, yielding a dense georeferenced camera pose for each frame.

² <https://github.com/iVS3D/iVS3D>

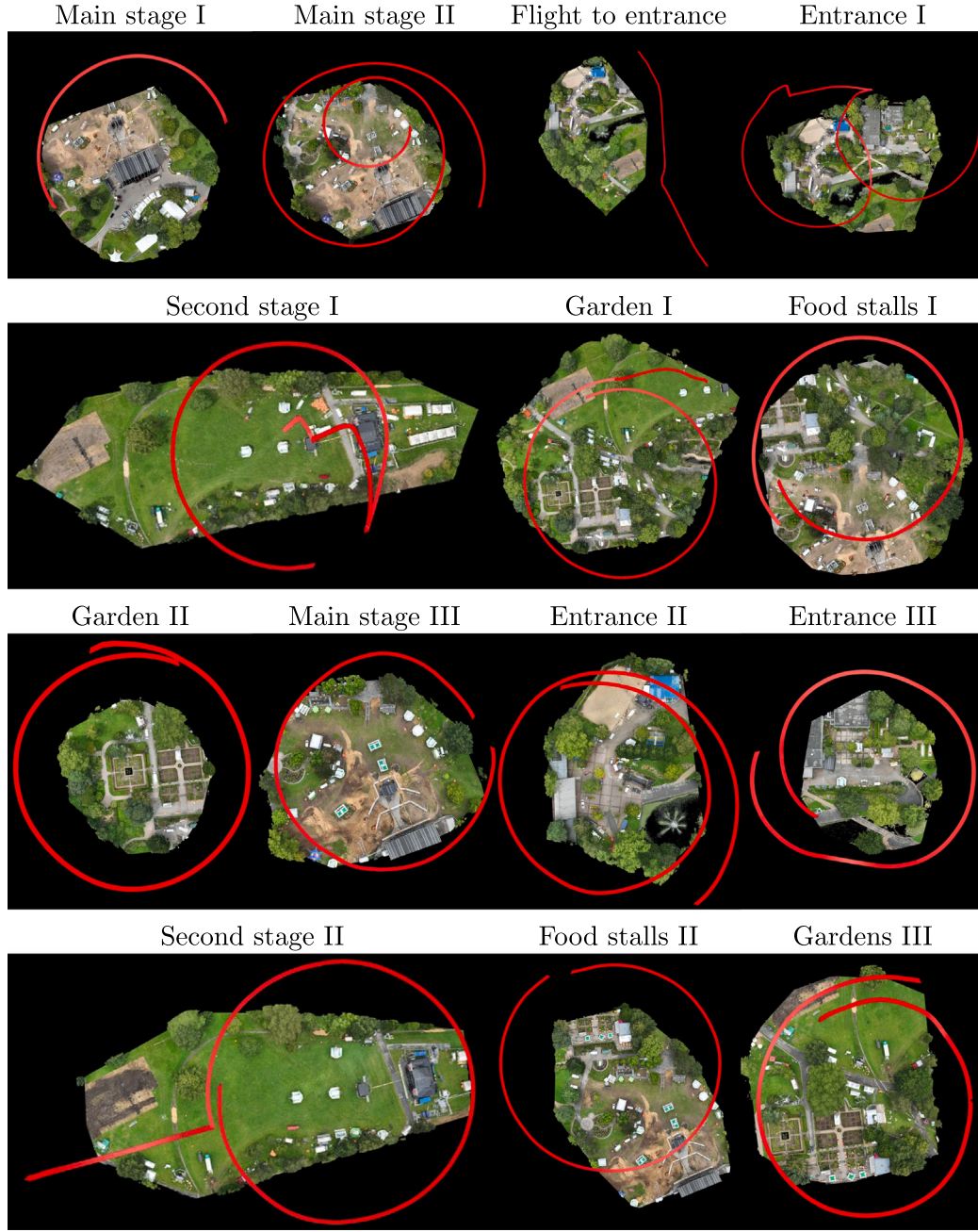


Fig. 2. Camera trajectories (red lines) of the 14 videos with the corresponding section of the ground truth. The manually defined polygons used to crop the LiDAR point clouds are primarily based on the area covered, the flight altitude and the camera angle. The videos are arranged chronologically; from scene *Main stage III* onwards, the flights took place on Day 2.

All camera trajectories start at the first video frame, but some trajectories end early to remove unnecessary images like the landing sequence. The still images from the double-grid flights are georeferenced with the same procedure. The scenes are reconstructed with COLMAP, transformed into the ENU frame using the per-image GNSS positions, and rigidly aligned to the reference LiDAR scan with ICP, so that the resulting camera poses are located in the same common coordinate system.

Although an RTK base station was used and the recorded GNSS positions are of high absolute accuracy, we observe a residual rigid offset of up to approximately 1 m between the SfM trajectory after the similarity transformation and the reference LiDAR scan. This offset does not reflect a limitation of the RTK positioning itself, but most likely results from a combination of systematic effects along the processing pipeline,

such as the unmodeled lever-arm between the RTK antenna and the camera, residual time-synchronization errors, small inaccuracies in the estimated similarity transformation, or slightly different systematic biases in the independent georeferencing of the LiDAR reference. As these effects appear predominantly as a global rigid shift rather than a scene-dependent deformation, they can be compensated by the rigid ICP alignment. After ICP, we no longer observe systematic offsets in our qualitative results that could be attributed to trajectory errors.

3.4. Setup of the evaluated approaches

We involve four MVS baselines: COLMAP (Schönberger et al., 2016) and ACMMP (Xu et al., 2022), which are conventional depth estimation methods; the learning-based MVSFormer++ (Cao et al., 2024); and



Fig. 3. Examples of challenging geometry. On the right side, the color-coded height map of the LiDAR scan is visualized.



Fig. 4. Dynamic scenes during data acquisition.



Fig. 5. RGB and thermal infrared still images from the double-grid flights.

PGSR, which leverages advanced scene representations via 3D Gaussian Splatting (3DGS) (Chen et al., 2025). For all methods, we use the full image resolution and otherwise default parameters. As camera poses, we use the 3 FPS trajectories, which means using every 10th image, starting with the first one.

The first baseline we use is COLMAP, as it is to this day a widely used open-source end-to-end SfM and MVS framework (Schönberger and Frahm, 2016; Schönberger et al., 2016). For SfM, it extracts Scale-Invariant Feature Transform (SIFT) features (Lowe, 2004), matches them and performs incremental reconstruction with periodic global bundle adjustment. Both extrinsic and intrinsic parameters can be jointly estimated. For MVS, COLMAP uses a probabilistic PatchMatch stereo algorithm extending the work of Zheng et al. (2014) to jointly

estimate depth and normal vectors per pixel, guided by view selection priorities and using photometric consistency with geometric checks. The per-pixel depth and normal maps estimated by the PatchMatch stereo algorithm are fused into a dense 3D point cloud by applying photometric and geometric consistency checks across multiple views, together with visibility constraints.

MVS methods that rely on photometric consistency during dense image matching often struggle in low-textured regions. ACMMP (Xu et al., 2022) addresses this by unifying ACMH/ACMM/ACMP (Xu and Tao, 2019, 2020) within one framework. In contrast to COLMAP's probabilistic view selection, ACMMP adopts adaptive checkerboard sampling and multi-hypothesis view selection to increase parallelism, enlarge the support region, and improve accuracy. A multi-scale, coarse-to-fine

strategy propagates reliable estimates from coarser to finer levels to reduce ambiguity in texture-poor areas. A planar prior further regularizes depth to improve completeness in low-textured regions. Similar to COLMAP, the estimated depth maps are combined into a consistent 3D point cloud by using depth reprojection geometric consistency filtering to discard unreliable points.

To tackle challenging geometries, more and more approaches are relying on Deep Neural Networks (DNNs) and thus a learning-based system to approach MVS. MVSFormer++ refines transformer models for MVS by using attention and positional encoding in each module (Cao et al., 2024). A frozen DINOv2 (Oquab et al., 2024) backbone is extended with a side view attention branch that incorporates cross-view information using linear attention. For cost volume regularization, a pure transformer model replaces commonly used 3D CNNs, applying global attention and introducing a frustocircular 3D positional encoding and adaptive attention scaling to handle varying sequence lengths and to mitigate attention dilution. Like the authors, we use fusibile³ as the fusion method presented by Galliani et al. (2015). We use the checkpoint provided by the authors for evaluation on the Tanks and Temples dataset (Knapitsch et al., 2017). Here, the DTU and Blended-MVS datasets are used as training data (Jensen et al., 2014; Yao et al., 2020).

While the other three approaches all estimate depth maps for each image and then fuse them into a point cloud, advanced scene representations such as NeRFs (Mildenhall et al., 2021) or 3DGS (Kerbl et al., 2023) use more complex representations of the entire 3D scene. However, the 3D geometry extracted from these is often unsatisfactory. PGSR (Chen et al., 2025) addresses this by flattening Gaussians along the smallest axis for regularization and ensuring geometric consistency via an additional MVS loss term. For comparability, we use the authors' approach by rendering depth maps from the 3DGS representation and integrating them into a Truncated Signed Distance Function (TSDF) volume with a voxel size of 5 cm.

3.5. Evaluation metrics

As evaluation metrics, we use the L1 absolute error and the root-mean-square error (RMSE) error to assess accuracy alongside the completeness to measure how well the reference data is covered. For completeness, a threshold value θ of 5 cm is chosen, as the images have a high resolution and comparatively low flight altitude, which results in a ground sampling distance (GSD) between 2 cm and 3 cm for the still images depending on the flight altitude. To make the results compatible with the DTU and Tanks and Temples datasets (Jensen et al., 2014; Knapitsch et al., 2017), the Chamfer distance and the F1-Score are also specified. Since point clouds are inherently unordered data structures, to calculate each metric for each point, the nearest corresponding point in the other point cloud is selected. Because the LiDAR scan is significantly larger than the areas covered by individual videos, we define a convex polygon for each video scene, which is used to crop the ground truth and estimate before calculating the metrics. Additionally to compare point clouds of varying density, we perform voxel downsampling with a voxel resolution of 2.5 cm, which is calculated similarly to Knapitsch et al. (2017) from half the threshold value θ used to compute the completeness as well as F1-Score. The code for the evaluation routine can be found here⁴. The metrics are formalized below:

$$\text{L1-abs}(\mathcal{P}_{est}, \mathcal{P}_{gt}) = \frac{1}{|\mathcal{E}|} \sum_{p \in \mathcal{E}} |\mathcal{P}_{est}(p) - \mathcal{P}_{gt}(p)| \quad (1)$$

$$\text{RMSE}(\mathcal{P}_{est}, \mathcal{P}_{gt}) = \sqrt{\frac{1}{|\mathcal{E}|} \sum_{p \in \mathcal{E}} (\mathcal{P}_{est}(p) - \mathcal{P}_{gt}(p))^2} \quad (2)$$

L1-abs($\mathcal{P}_{est}, \mathcal{P}_{gt}$) and RMSE($\mathcal{P}_{est}, \mathcal{P}_{gt}$) specify the distance between each point p in the estimate $\mathcal{P}_{est}$ to its nearest neighboring point in the ground truth point cloud $\mathcal{P}_{gt}$ averaged by the number of points $\mathcal{E}$ in the estimate. To exclude outliers, a threshold value of 90 % is used to exclude the worst 10 % for L1-abs and RMSE, like described by Seitz et al. (2006).

$$\text{Precision}(\mathcal{P}_{est}, \mathcal{P}_{gt}|\theta) = \frac{100}{|\mathcal{E}|} \sum_{p \in \mathcal{E}} |\mathcal{P}_{est}(p) - \mathcal{P}_{gt}(p)| < \theta \quad (3)$$

$$\text{Recall}(\mathcal{P}_{est}, \mathcal{P}_{gt}|\theta) = \frac{100}{|\mathcal{G}|} \sum_{p \in \mathcal{G}} |\mathcal{P}_{est}(p) - \mathcal{P}_{gt}(p)| < \theta \quad (4)$$

$$\text{F1-Score}(\mathcal{P}_{est}, \mathcal{P}_{gt}|\theta) = \frac{2\text{Recall}(\mathcal{P}_{est}, \mathcal{P}_{gt}|\theta) \text{Precision}(\mathcal{P}_{est}, \mathcal{P}_{gt}|\theta)}{\text{Recall}(\mathcal{P}_{est}, \mathcal{P}_{gt}|\theta) + \text{Precision}(\mathcal{P}_{est}, \mathcal{P}_{gt}|\theta)} \quad (5)$$

The F1-Score represents the harmonic mean between recall and precision. Analogous to the definition in Knapitsch et al. (2017), we calculate for each point in the estimate $\mathcal{P}_{est}$ how far it is from a corresponding point in the ground truth point cloud $\mathcal{P}_{gt}$ and whether this distance is less than θ . This ratio multiplied by 100 represents the precision metric. Similarly, for recall, starting from the ground truth point cloud $\mathcal{P}_{gt}$, we calculate how many points are within reach of θ to the estimated point cloud $\mathcal{P}_{est}$. The completeness metric is equivalent to the recall but is usually represented in percent instead of a score from 0 to 100.

$$\text{Chamfer}(\mathcal{P}_{est}, \mathcal{P}_{gt}) = \frac{1}{|\mathcal{E}|} \sum_{p \in \mathcal{E}} |\mathcal{P}_{est}(p) - \mathcal{P}_{gt}(p)| + \frac{1}{|\mathcal{G}|} \sum_{p \in \mathcal{G}} |\mathcal{P}_{gt}(p) - \mathcal{P}_{est}(p)| \quad (6)$$

Another widely used metric combining both a measure for completeness and accuracy in one score is the Chamfer distance, which represents the symmetrical distance between the point clouds $\mathcal{P}_{est}$ and $\mathcal{P}_{gt}$.

4. Experimental results

In the following section, we evaluate the four baseline approaches by comparing the reconstructed point clouds with the LiDAR ground truth. Across all scenes, we use voxel downsampling with a resolution of 2.5 cm and a threshold value θ of 5 cm for completeness and F1-Score. The MVSFormer++ produces extremely large point clouds, which prevents evaluation on some large-scale scenes due to implementation limitations. In these cases, we reduce the point clouds to 3 billion points before cropping with the provided polygons. Since the results of PGSR are Gaussian splats, we first extract a mesh by rendering depth maps and projecting them into a TSDF voxel volume. To compare the results with the other methods, we then sample a point cloud from the surface. For comparability the number of points is based on the number of points reconstructed by ACMMP. This point cloud is then cropped using the defined polygon. Tables 4 and 5 show the quantitative results for each sequence split per day of the measurement campaign. Figs. 6 and 7 show qualitative results.

Although the individual scenes differ in terms of flight trajectories and content, trends in the performance of the methods can be identified. Especially in the case of the symmetric metrics of Chamfer distance and F1-Score, which take completeness and accuracy into account equally, the MVSFormer++ stands out and delivers better results than all other methods on most scenes, as can be seen in Tables 4 and 5. The completeness, which also contributes to the F1-Score as recall, is significantly higher using the MVSFormer++ across almost all scenes, which might explain these results. However, if a closer examination of the values for L1-abs and RMSE is conducted, it becomes apparent that the conventional methods COLMAP and ACMMP deliver more accurate results for every scene. Since the percentage difference of the RMSE is higher than that of L1-abs, it is reasonable to assume that the reconstructed point clouds of MVSFormer++ contain more outliers than the conventional methods. In addition, ACMMP consistently performs better in terms of L1-abs and RMSE than COLMAP across almost all scenes.

³ <https://github.com/YoYo000/fusibile>

⁴ <https://github.com/JB3D-Dataset/Multi-View-Stereo>

Table 4

Quantitative results of the reconstructed point clouds from Day 1. The results are ordered by video sequence, with the second-best result per column and scene highlighted underlined and the best result additionally by bold print. The number of points is listed after cropping but before voxel downsampling is applied. Cpl. denotes completeness and CD denotes Chamfer distance.

Scene	Method	Cpl. ↑	L1-abs ↓	RMSE ↓	CD ↓	F1-Score ↑	No. Points	Runtime ↓
			in m	in m	in m	in percent		in min
Main stage I	COLMAP	0.3749	<u>0.0877</u>	<u>0.1131</u>	<u>0.4875</u>	36.50	30,442,580	191
	ACMMP	0.4598	0.0757	0.0953	0.4970	42.80	105,904,717	151
	PGSR	0.2991	0.1271	0.1771	0.5876	29.74	104,186,194	320
	MVSF++	<u>0.6622</u>	0.0969	0.1195	0.4320	<u>39.26</u>	1,227,616,695	<u>180</u>
Main stage II	COLMAP	0.5175	<u>0.0759</u>	<u>0.0997</u>	0.2986	<u>47.21</u>	32,433,389	265
	ACMMP	<u>0.6219</u>	0.0641	0.0807	<u>0.2809</u>	54.51	116,476,716	<u>169</u>
	PGSR	0.4557	0.1029	0.1487	0.4026	42.11	58,215,170	329
	MVSF++	0.8374	0.1003	0.1282	0.2689	42.76	1,317,229,956	101
Flight to entrance	COLMAP	0.0940	0.0978	0.1251	0.8916	14.38	11,843,414	115
	ACMMP	<u>0.1238</u>	<u>0.1056</u>	<u>0.1380</u>	1.0395	<u>17.49</u>	24,720,215	<u>69</u>
	PGSR	0.1024	0.1862	0.2675	1.4728	14.01	23,635,190	284
	MVSF++	0.3583	0.1083	0.1394	0.7607	31.16	337,194,622	31
Entrance I	COLMAP	0.4090	<u>0.0890</u>	<u>0.1187</u>	0.6081	38.72	80,629,251	601
	ACMMP	0.4562	0.0847	0.1092	0.6039	40.16	182,198,498	411
	PGSR	0.2738	0.2222	0.4353	1.3218	28.83	141,569,317	473
	MVSF++	0.6925	0.0954	0.1242	0.5336	44.15	1,491,293,674	185
Second stage I	COLMAP	0.3750	<u>0.0465</u>	<u>0.0550</u>	<u>0.3417</u>	45.23	28,656,167	228
	ACMMP	0.4740	0.0437	0.0516	0.3495	53.01	89,421,442	160
	PGSR	0.3662	0.0707	0.0956	0.4835	41.01	60,505,410	242
	MVSF++	0.7027	0.0608	0.0721	0.2972	54.06	1,049,621,579	118
Garden I	COLMAP	0.3035	<u>0.0553</u>	<u>0.0663</u>	<u>0.4297</u>	37.83	59,891,846	364
	ACMMP	<u>0.3761</u>	0.0532	0.0631	0.4659	<u>43.39</u>	147,043,957	<u>273</u>
	PGSR	0.2962	0.0840	0.1151	0.6390	34.69	134,458,919	335
	MVSF++	0.6600	0.0683	0.0817	0.3647	49.52	1,763,536,880	122
Food stalls I	COLMAP	0.3668	<u>0.0617</u>	<u>0.0739</u>	<u>0.3158</u>	40.38	47,876,793	269
	ACMMP	<u>0.4363</u>	0.0589	0.0696	0.3389	<u>45.00</u>	120,689,162	<u>177</u>
	PGSR	0.3029	0.0988	0.1338	0.4565	32.56	128,692,188	332
	MVSF++	0.7048	0.0760	0.0916	0.2745	46.99	1,473,521,911	91
Garden II	COLMAP	0.4690	0.0399	0.0453	<u>0.2554</u>	54.26	23,169,515	155
	ACMMP	<u>0.5062</u>	<u>0.0402</u>	<u>0.0458</u>	0.2580	56.44	60,818,422	102
	PGSR	0.3481	0.0463	0.0540	0.3195	43.41	57,297,885	295
	MVSF++	0.7430	0.0468	0.0539	0.2114	63.90	748,753,837	48

Table 5

Quantitative results of the reconstructed point clouds from Day 2. The results are ordered by video sequence, with the second-best result per column and scene highlighted underlined and the best result additionally by bold print. The number of points is listed after cropping but before voxel downsampling is applied. Cpl. denotes completeness and CD denotes Chamfer distance.

Scene	Method	Cpl. ↑	L1-abs ↓	RMSE ↓	CD ↓	F1-Score ↑	No. Points	Runtime ↓
			in m	in m	in m	in percent		in min
Main stage III	COLMAP	0.5451	<u>0.0470</u>	<u>0.0599</u>	<u>0.2177</u>	57.08	42,993,775	295
	ACMMP	<u>0.6317</u>	0.0420	0.0520	0.2307	63.07	165,977,111	<u>220</u>
	PGSR	0.4819	0.0699	0.1014	0.3430	50.02	104,209,140	290
	MVSF++	0.7933	0.0634	0.0778	0.1936	<u>57.14</u>	1,866,996,966	115
Entrance II	COLMAP	0.3962	0.0455	0.0542	<u>0.3808</u>	47.49	43,033,958	287
	ACMMP	<u>0.4435</u>	<u>0.0481</u>	<u>0.0575</u>	0.3830	<u>49.77</u>	111,911,799	210
	PGSR	0.3075	0.0923	0.1422	0.7293	36.30	64,841,699	22
	MVSF++	0.6807	0.0642	0.0785	0.3215	53.61	1,296,364,042	<u>94</u>
Entrance III	COLMAP	0.6056	0.0364	0.0430	0.1726	64.42	46,578,574	270
	ACMMP	<u>0.6311</u>	<u>0.0386</u>	<u>0.0460</u>	0.1800	64.55	133,270,322	201
	PGSR	0.4479	0.0631	0.0924	0.4569	49.18	154,807,660	18
	MVSF++	0.8061	0.0502	0.0602	<u>0.1732</u>	64.95	1,487,794,705	<u>95</u>
Second stage III	COLMAP	0.4594	<u>0.0498</u>	<u>0.0633</u>	<u>0.2257</u>	50.97	53,761,062	428
	ACMMP	<u>0.5768</u>	0.0468	0.0590	0.2336	58.39	184,737,038	319
	PGSR	0.3338	0.0751	0.1035	0.4670	38.41	105,417,681	30
	MVSF++	0.7670	0.0678	0.0858	0.2208	<u>55.91</u>	1,659,169,399	<u>146</u>
Food stalls II	COLMAP	0.3318	<u>0.0438</u>	<u>0.0529</u>	0.2662	43.03	21,120,005	142
	ACMMP	<u>0.4357</u>	0.0404	0.0485	<u>0.2527</u>	<u>52.01</u>	58,025,449	102
	PGSR	0.2890	0.0690	0.0899	0.3627	35.42	35,938,100	15
	MVSF++	0.7366	0.0558	0.0678	0.1915	59.75	695,836,374	<u>50</u>
Garden III	COLMAP	0.3188	<u>0.0362</u>	<u>0.0415</u>	0.3793	43.61	21,302,682	143
	ACMMP	<u>0.4258</u>	0.0346	0.0401	<u>0.3601</u>	<u>53.17</u>	50,513,244	103
	PGSR	0.3075	0.0523	0.0666	0.4636	39.60	36,144,080	12
	MVSF++	0.6481	0.0458	0.0535	0.3122	60.83	691,210,743	<u>46</u>

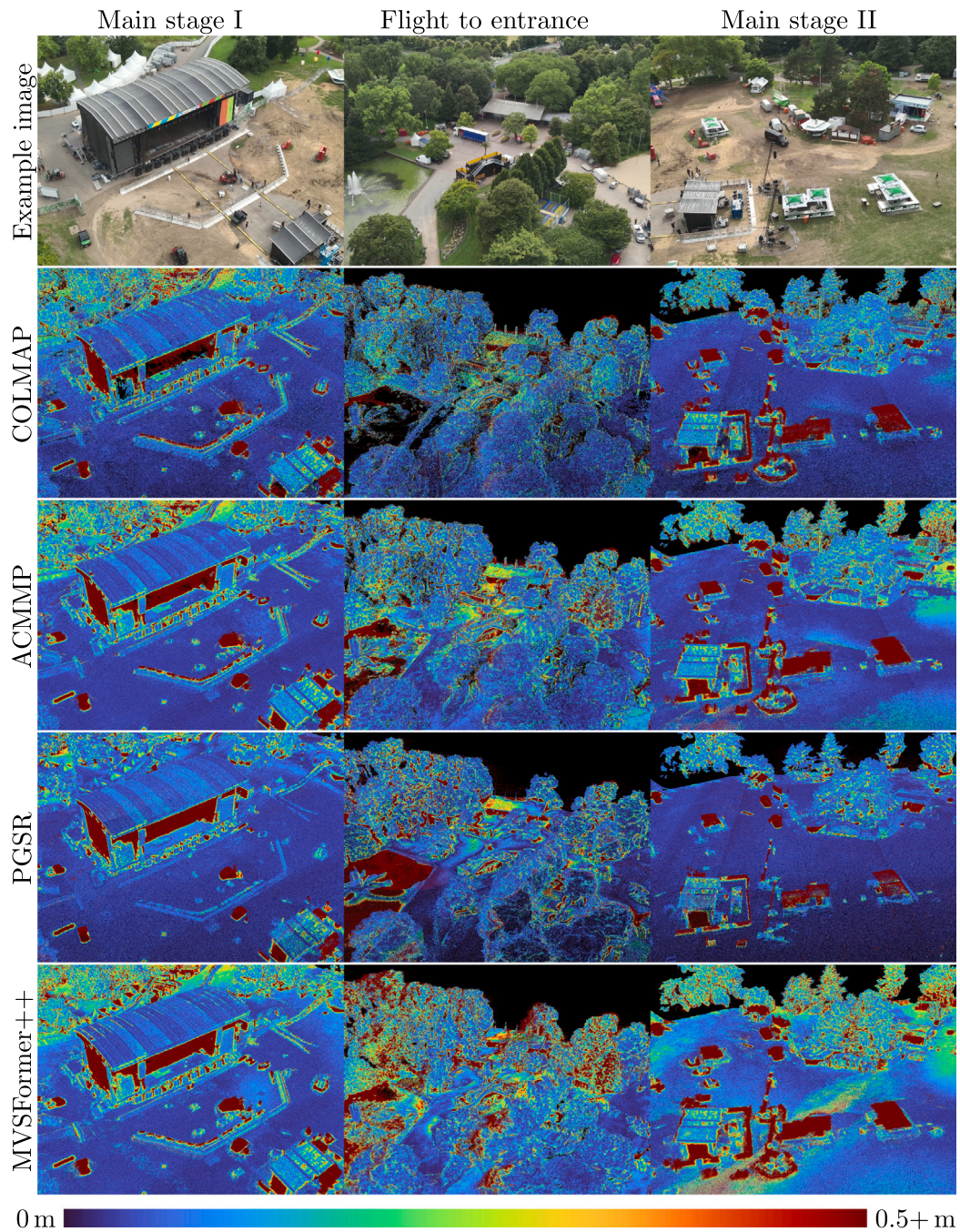


Fig. 6. Qualitative results of the reconstructed point clouds. The first row shows exemplary input images. The following rows display the results from the different methods rendered from the same perspective, with the error color coded.

In contrast, the results of PGSR fall short of those achieved by the other methods. However, despite the fact that PGSR is not primarily intended for 3D reconstruction, the gap is relatively small. In particular, integrating the rendered depth maps into a voxel volume with subsequent mesh extraction and point cloud sampling could distort the results favoring planar areas. As can be seen in the qualitative results of Figs. 6 and 7, a positive side effect of this approach could be that planar areas such as the ground level seem to be less error prone. This effect is particularly visible in the right column in Fig. 7. Even fine details, such as the radio mast visible in the middle column, can be reconstructed well by most methods. However, MVSFormer++ has difficulties with this, which may be related to the use of pre-trained weights and the lack of additional fine-tuning on our side. Not surprisingly in both

Fig. 6 and Fig. 7 water areas are highly inaccurate in all experiments, as not only do the 3D reconstructions fail, but the LiDAR scan is also likely to be of poor quality there.

All experiments were conducted on one Nvidia L40, with the run time of all steps measured end-to-end. In the case of COLMAP, ACMMP, and MVSFormer++, this includes the fusion into a point cloud. For PGSR, it includes the rendering of the depth maps, TSDF volume integration and the extraction of the triangle meshes. When considering the run time, it should be noted that, unlike the other methods, PGSR does not have a clearly defined endpoint. Instead, 30,000 iterations are used, as in the default configuration. As can be seen from Tables 4 and 5, the MVSFormer++ is particularly fast in producing results, which is noteworthy given that it delivers very dense point clouds.

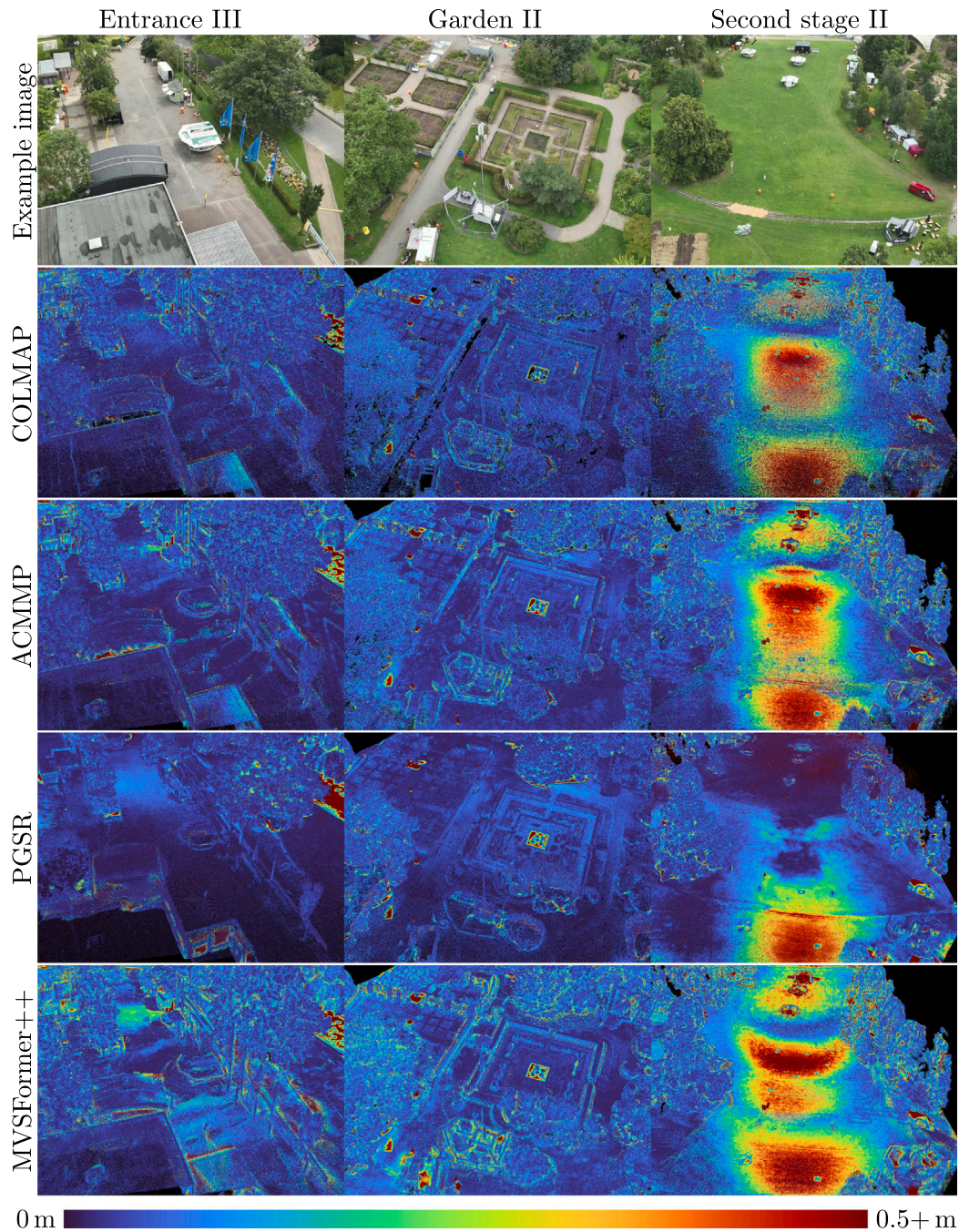


Fig. 7. Qualitative results of the reconstructed point clouds. The first row shows exemplary input images. The following rows display the results from the different methods rendered from the same perspective, with the error color coded.

5. Discussion

As can be seen in the previous section, there are clear differences between the evaluated approaches. Although MVSFormer++ delivers the best F1-Score, this is in large part due to the much denser and more complete point clouds. If we look only at accuracy in the form of L1-abs and RMSE error, the conventional methods, and ACMMP in particular, perform significantly better. However, it should be noted that the standard parameterization is used for all methods. Particularly given the very large number of points resulting from MVSFormer++, it would probably be reasonable to increase the confidence threshold for the depth maps. The performance of MVSFormer++ is particularly

noteworthy, as we are using the weights and parameters from the Tanks and Temples configuration and only increasing the image resolution.

Improving the parameterization could also lead to significantly better results with PGSR, as the standard fusion approach integrates the depth maps into a TSDF volume and then extracts a triangle mesh. This step is sensitive, as the parameters for the voxel volume determine which details can be reconstructed at all. A different type of fusion or a direct evaluation of the depth maps could deliver better results. Despite the huge point clouds, MVSFormer++ is the method with the lowest runtime across all scenes. PGSR is faster in some scenes, but this is probably because 30,000 iterations are used by default, regardless of the number of images or the dimensions of the scene. Especially



Fig. 8. Novel view synthesis of challenging objects using PGSR. Both images were rendered from camera angles that differ from the video frames used for the reconstruction.



Fig. 9. Changes within the scene. On the left and center are two LiDAR scans from Day 1 and Day 2 displayed. On the right side, the changes are color coded. Red highlights major changes, while blue symbolizes no significant changes.

in larger scenes, more iterations could lead to better results, while in smaller scenes, early stopping could lead to shorter run times. While PGSR does not produce the best results in terms of geometric quality, the Gaussian splatting output is visually very appealing, as can be seen in novel view synthesis experiments shown in Fig. 8. The photorealistic representation of fine details such as grids, ladders, and vegetation is a major advantage over the other approaches.

While we only examined 3D reconstruction in Section 4, the dataset also opens the possibility of detecting changes over time, as LiDAR scans, videos, and still images are all referenced in one coordinate system. In addition, the measurement campaign spans two days, and all areas are covered on both days. A simple method of annotating changes is the comparison of the LiDAR scans from the two days with each other. Fig. 9 shows this by displaying renderings of the colored LiDAR scans from Day 1 and Day 2. Obvious changes, such as the newly arrived trucks at the bottom center or the additional food stalls, appear in red in the point cloud in the right column.

While we have only evaluated the reconstructed point clouds in this work, the LiDAR point clouds can also be projected onto the image plane using the provided camera poses to render artificial ground truth depth maps. A corresponding script for this can be found here: ⁵. Since the dataset includes 30 FPS camera trajectories in addition to the LiDAR point clouds, SfM methods could also be evaluated. However, there is a limitation here in that these trajectories were also created photogrammetrically without additional tools like laser range finder or Terrestrial Laser Scanner (TLS). Another limitation of the dataset is the dynamic elements such as vehicles and people, which can be challenging for 3D reconstruction methods. But such dynamic objects are of course interesting for other downstream tasks.

As evident from the quantitative results in Tables 4 and 5, the choice of evaluation metric significantly affects the outcome, since different considerations underlie each metric. By evaluating only the best 90 % of points for L1-abs and RMSE, we mainly measure inlier

accuracy with these metrics. In contrast, the Chamfer distance evaluates all points, including outliers. To assess the completeness of a reconstructed model, the coverage of the ground truth point cloud is measured. If a point cloud is filtered more aggressively, the L1-abs/RMSE is usually lower, but the completeness decreases at the same time. Ideally, a better methodology would have a lower error rate and a higher degree of completeness. Surprisingly, COLMAP achieves a lower Chamfer distance than ACMMP, despite ACMMP having both a lower L1-abs/RMSE error and a higher completeness value. This could be due to the symmetrical weighting of the Chamfer distance and a different distribution of outliers, which may have a heavier tail of bad points. Similarly, the F1-Score equally weighs completeness/recall and precision, resulting in the MVSFormer++ achieving very good results due to its high completeness despite noticeably worse L1-abs/RMSE values. The higher error can also be seen in the qualitative results in the last row of Figs. 6 and 7. While it is advantageous that the Chamfer distance and F1-Score represent a combined value that simplifies method comparison, many use cases likely require different weighting, which is facilitated by splitting them into two metrics. For example, tasks such as land surveying may prioritize accuracy, while visualization applications require a more complete reconstruction.

6. Conclusion

In conclusion, we present the first aerial dataset that includes videos, LiDAR point clouds and georeferenced 30 FPS camera trajectories for benchmarking MVS methods. In addition, we provide the employed evaluation routines ⁶. As a baseline, we examined both conventional methods in the form of COLMAP and ACMMP, as well as learning-based methods such as MVSFormer++, and PGSR relying on an advanced scene representation. Although PGSR performs slightly worse than the rest, it is primarily intended for creating photorealistic models using Gaussian splatting. MVSFormer++ achieves the highest F1-Score and completeness across all scenes with a comparatively

⁵ <https://github.com/JB3D-Dataset/Multi-View-Stereo>

⁶ <https://github.com/JB3D-Dataset/Multi-View-Stereo>

low runtime. However, ACMMP achieves significantly lower errors for L1-abs and RMSE.

We also provide 5000 high-resolution still images that were used to map the area, 3000 of which have simultaneously captured thermal infrared images. In addition, we provided five other videos with registered camera trajectories that are either too challenging for MVS or lie outside the LiDAR coverage. Since the dataset was recorded over two days, changes are visible in the scene, which opens the possibility of performing change detection. This is further facilitated by the fact that all videos, still images, and LiDAR scans are referenced in one single joint coordinate system.

CRedit authorship contribution statement

M. Hermann: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Data curation, Conceptualization. **D. Böhmer:** Software, Resources, Investigation, Data curation. **S. Hinz:** Writing – review & editing, Supervision, Funding acquisition. **M. Weinmann:** Writing – review & editing, Writing – original draft, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We gratefully acknowledge the support and cooperation of the festival organizers, whose efforts made this measurement campaign possible.

We acknowledge support by the KIT-Publication Fund of the Karlsruhe Institute of Technology.

References

- Aanaes, H., Jensen, R.R., Vogiatzis, G., Tola, E., Dahl, A.B., 2016. Large-scale data for multiple-view stereopsis. *Int. J. Comput. Vis.* 120, 153–168. <http://dx.doi.org/10.1007/s11263-016-0902-9>.
- Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P.P., 2021. Mip-NeRF: A multiscale representation for anti-aliasing neural radiance fields. In: *Proc. IEEE/CVF International Conference on Computer Vision*. pp. 5835–5844.
- Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P., 2022. Mip-NeRF 360: Unbounded anti-aliased neural radiance fields. In: *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 5460–5469.
- Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P., 2023. Zip-NeRF: Anti-aliased grid-based neural radiance fields. In: *Proc. IEEE/CVF International Conference on Computer Vision*. pp. 19697–19705.
- Besl, P.J., McKay, N.D., 1992. Method for registration of 3-D shapes. In: *Proc. Sensor Fusion IV: Control Paradigms and Data Structures*, vol. 1611, SPIE, pp. 586–606.
- Buschmann, B., Dogaru, A., Eisemann, E., Weinmann, M., Egger, B., 2024. RANRAC: Robust neural scene representations via random ray consensus. In: *Proc. European Conference on Computer Vision*. pp. 126–143.
- Cao, C., Ren, X., Fu, Y., 2022. MVSFormer: Multi-view stereo by learning robust image features and temperature-based depth. *Trans. Mach. Learn. Res.* URL <https://openreview.net/forum?id=2VVR6JfwNo>.
- Cao, C., Ren, X., Fu, Y., 2024. MVSFormer++: Revealing the devil in transformer's details for multi-view stereo. In: *Proc. International Conference on Learning Representations*.
- Chen, K., Dong, B., Wang, Z., Cheng, P., Yan, M., Sun, X., Weinmann, M., Weinmann, M., 2024. PriNeRF: Prior constrained neural radiance field for robust novel view synthesis of urban scenes with fewer views. *ISPRS J. Photogramm. Remote Sens.* 215, 383–399.
- Chen, D., Li, H., Ye, W., Wang, Y., Xie, W., Zhai, S., Wang, N., Liu, H., Bao, H., Zhang, G., 2025. PGSR: Planar-Based Gaussian Splatting for Efficient and High-Fidelity Surface Reconstruction. *IEEE Trans. Vis. Comput. Graphics* 31 (09), 6100–6111. <http://dx.doi.org/10.1109/TVCG.2024.3494046>, URL <https://doi.ieeecomputersociety.org/10.1109/TVCG.2024.3494046>.
- Chen, A., Xu, Z., Geiger, A., Yu, J., Su, H., 2022. TensorRF: Tensorial radiance fields. In: *Proc. European Conference on Computer Vision*. pp. 333–350.
- Ding, Y., Yuan, W., Zhu, Q., Zhang, H., Liu, X., Wang, Y., Liu, X., 2022. TransMVSNet: Global context-aware multi-view stereo network with transformers. In: *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 8575–8584.
- Furukawa, Y., Ponce, J., 2009. Accurate, dense, and robust multiview stereopsis. *IEEE Trans. Pattern Anal. Mach. Intell.* 32 (8), 1362–1376.
- Galliani, S., Lasinger, K., Schindler, K., 2015. Massively parallel multiview stereopsis by surface normal diffusion. In: *Proc. IEEE International Conference on Computer Vision*. pp. 873–881.
- Hermann, M., Kwak, H., Ruf, B., Weinmann, M., 2024a. Leveraging neural radiance fields for large-scale 3D reconstruction from aerial imagery. *Remote. Sens.* 16 (24).
- Hermann, M., Pollok, T., Brommer, D., Zahn, D., 2021. IVS3D: An open source framework for intelligent video sampling and preprocessing to facilitate 3D reconstruction. In: *Proc. International Symposium on Visual Computing*. Springer-Verlag, Berlin, Heidelberg, pp. 441–454. http://dx.doi.org/10.1007/978-3-030-90439-5_35.
- Hermann, M., Weinmann, M., Nex, F., Stathopoulou, E., Remondino, F., Jutzi, B., Ruf, B., 2024b. Depth estimation and 3D reconstruction from UAV-borne imagery: Evaluation on the UseGeo dataset. *ISPRS Open J. Photogramm. Remote. Sens.* 100065. <http://dx.doi.org/10.1016/j.ophoto.2024.100065>, URL <https://www.sciencedirect.com/science/article/pii/S2667393224000085>.
- Huang, P.H., Matzen, K., Kopf, J., Ahuja, N., Huang, J.B., 2018. DeepMVS: Learning multi-view stereopsis. In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition*. pp. 2821–2830.
- Jensen, R., Dahl, A., Vogiatzis, G., Tola, E., Aanaes, H., 2014. Large scale multi-view stereopsis evaluation. In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, pp. 406–413.
- Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G., 2023. 3D Gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.* 42 (4), 139:1–139:14.
- Knapitsch, A., Park, J., Zhou, Q.Y., Koltun, V., 2017. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Trans. Graph.* 36 (4).
- Lin, C.H., Ma, W.C., Torralba, A., Lucey, S., 2021. BARF: Bundle-adjusting neural radiance fields. In: *Proc. IEEE/CVF International Conference on Computer Vision*. pp. 5721–5731.
- Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* 60 (2), 91–110.
- Martin-Brualla, R., Radwan, N., Sajjadi, M.S.M., Barron, J.T., Dosovitskiy, A., Duckworth, D., 2021. NeRF in the Wild: Neural radiance fields for unconstrained photo collections. In: *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 7206–7215.
- Mi, Z., Xu, D., 2023. Switch-NeRF: Learning scene decomposition with mixture of experts for large-scale neural radiance fields. In: *Proc. 11th International Conference on Learning Representations*. pp. 1–15.
- Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R., 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM* 65 (1), 99–106.
- Müller, T., Evans, A., Schied, C., Keller, A., 2022. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.* 41 (4), 102:1–102:15.
- Nex, F., Gerke, M., Remondino, F., Przybilla, H.J., Baumker, M., Zurhorst, A., 2015. ISPRS benchmark for multi-platform photogrammetry. *ISPRS Ann. Photogramm. Remote. Sens. Spat. Inf. Sci.* II-3/W4, 135–142. <http://dx.doi.org/10.5194/isprsannals-II-3-W4-135-2015>, URL <https://isprs-annals.copernicus.org/articles/II-3-W4/135/2015/>.
- Nex, F., Stathopoulou, E., Remondino, F., Yang, M., Madhuanand, L., Yogender, Y., Alsadik, B., Weinmann, M., Jutzi, B., Qin, R., 2024. UseGeo – A UAV-based multi-sensor dataset for geospatial research. *ISPRS Open J. Photogramm. Remote. Sens.* 13, 100070. <http://dx.doi.org/10.1016/j.ophoto.2024.100070>, URL <https://www.sciencedirect.com/science/article/pii/S2667393224000140>.
- Niemeyer, M., Barron, J.T., Mildenhall, B., Sajjadi, M.S.M., Geiger, A., Radwan, N., 2022. RegNeRF: Regularizing neural radiance fields for view synthesis from sparse inputs. In: *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 5470–5480.
- Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P., 2024. DINOv2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.* URL <https://openreview.net/forum?id=a68SUt6zFt>.
- Peng, R., Wang, R., Wang, Z., Lai, Y., Wang, R., 2022. Rethinking depth estimation for multi-view stereo: A unified representation. In: *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 8635–8644.
- Ruano, S., Smolic, A., 2021. A benchmark for 3d reconstruction from aerial imagery in an urban environment. In: *Proc. 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, vol. 5 (VISAPP), pp. 732–741.
- Sabour, S., Vora, S., Duckworth, D., Krasin, I., Fleet, D.J., Tagliasacchi, A., 2023. RobustNeRF: Ignoring distractors with robust losses. In: *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 20626–20636.
- Scharstein, D., Szeliski, R., 2002. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int. J. Comput. Vis.* 47 (1), 7–42.
- Schönberger, J.L., Frahm, J.M., 2016. Structure-from-motion revisited. In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition*. pp. 4104–4113.

- Schönberger, J.L., Zheng, E., Pollefeys, M., Frahm, J.M., 2016. Pixelwise view selection for unstructured multi-view stereo. In: Proc. European Conference on Computer Vision. pp. 501–518.
- Schöps, T., Schönberger, J.L., Galliani, S., Sattler, T., Schindler, K., Pollefeys, M., Geiger, A., 2017. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition. pp. 2538–2547. <http://dx.doi.org/10.1109/CVPR.2017.272>.
- Seitz, S., Curless, B., Diebel, J., Scharstein, D., Szeliski, R., 2006. A comparison and evaluation of multi-view stereo reconstruction algorithms. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 519–528. <http://dx.doi.org/10.1109/CVPR.2006.19>.
- Snively, N., Seitz, S.M., Szeliski, R., 2006. Photo Tourism: Exploring photo collections in 3D. In: Proc. ACM Siggraph. pp. 835–846.
- Stathopoulou, E.K., Remondino, F., 2023. A survey on conventional and learning-based methods for multi-view stereo. *Photogramm. Rec.* 38 (138), 374–407.
- Stathopoulou, E.K., Welponer, M., Remondino, F., 2019. Open-source image-based 3d reconstruction pipelines: Review, comparison and evaluation. *XLII-2/W17*, pp. 331–338.
- Strecha, C., von Hansen, W., Van Gool, L., Fua, P., Thoennessen, U., 2008. On benchmarking camera calibration and multi-view stereo for high resolution imagery. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–8. <http://dx.doi.org/10.1109/CVPR.2008.4587706>.
- Tancik, M., Casser, V., Yan, X., Pradhan, S., Mildenhall, B.P., Srinivasan, P., Barron, J.T., Kretschmar, H., 2022. Block-NeRF: Scalable large scene neural view synthesis. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8238–8248.
- Tewari, A., Thies, J., Mildenhall, B., Srinivasan, P., Treitschk, E., Yifan, W., Lassner, C., Sitzmann, V., Martin-Brualla, R., Lombardi, S., Simon, T., Theobalt, C., Nießner, M., Barron, J.T., Wetzstein, G., Zollhöfer, M., Golyanik, V., 2022. Advances in neural rendering. *Comput. Graph. Forum* 41 (2), 703–735.
- Turki, H., Ramanan, D., Satyanarayanan, M., 2022. Mega-NeRF: Scalable construction of large-scale NeRFs for virtual fly-throughs. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12912–12921.
- Wang, X., Wang, C., Liu, B., Zhou, X., Zhang, L., Zheng, J., Bai, X., 2021. Multi-view stereo in the deep learning era: A comprehensive review. *Displays* 70, 102102:1–102102:12.
- Xiangli, Y., Xu, L., Pan, X., Zhao, N., Rao, A., Theobalt, C., Dai, B., Lin, D., 2022. BungeeNeRF: Progressive neural radiance field for extreme multi-scale scene rendering. In: Proc. European Conference on Computer Vision. pp. 106–122.
- Xie, S., Zhang, L., Jeon, G., Yang, X., 2023. Remote sensing neural radiance fields for multi-view satellite photogrammetry. *Remote. Sens.* 15 (15), 3808:1–3808:17.
- Xu, Q., Kong, W., Tao, W., Pollefeys, M., 2022. Multi-scale geometric consistency guided and planar prior assisted multi-view stereo. *IEEE Trans. Pattern Anal. Mach. Intell.* 45 (4), 4945–4963.
- Xu, N., Qin, R., Huang, D., Remondino, F., 2024. Multi-tiling neural radiance field (NeRF)—geometric assessment on large-scale aerial datasets. *Photogramm. Rec.* 39 (188), 718–740.
- Xu, Q., Tao, W., 2019. Multi-scale geometric consistency guided multi-view stereo. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition. pp. 5483–5492.
- Xu, Q., Tao, W., 2020. Planar prior assisted PatchMatch multi-view stereo. In: Proc. AAAI Conference on Artificial Intelligence. pp. 12516–12523.
- Yan, Z., Mazzacca, G., Rigon, S., Farella, E.M., Trybala, P., Remondino, F., 2023. NeRFbK: a holistic dataset for benchmarking NeRF-based 3D reconstruction. *Int. Arch. Photogramm. Remote. Sens. Spat. Inf. Sci.* XLVIII-1/W3-2023, 219–226. <http://dx.doi.org/10.5194/isprs-archives-XLVIII-1-W3-2023-219-2023>, URL <https://isprs-archives.copernicus.org/articles/XLVIII-1-W3-2023/219/2023/>.
- Yao, Y., Luo, Z., Li, S., Fang, T., Quan, L., 2018. MVSNet: Depth inference for unstructured multi-view stereo. In: Proc. European Conference on Computer Vision. pp. 767–783.
- Yao, Y., Luo, Z., Li, S., Zhang, J., Ren, Y., Zhou, L., Fang, T., Quan, L., 2020. BlendedMVS: A large-scale dataset for generalized multi-view stereo networks. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition. pp. 1787–1796. <http://dx.doi.org/10.1109/CVPR42600.2020.00186>.
- Yu, A., Ye, V., Tancik, M., Kanazawa, A., 2021. pixelNeRF: Neural radiance fields from one or few images. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4576–4585.
- Zhang, Z., Peng, R., Hu, Y., Wang, R., 2023. GeoMVSNet: Learning multi-view stereo with geometry perception. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21508–21518.
- Zheng, E., Dunn, E., Jovic, V., Frahm, J.M., 2014. PatchMatch based joint view selection and depthmap estimation. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition. pp. 1510–1517.