POLITECNICO DI TORINO
Repository ISTITUZIONALE

UseGeo - A UAV-based multi-sensor dataset for geospatial research

Original
UseGeo - A UAV-based multi-sensor dataset for geospatial research / Nex, F.; Stathopoulou, E. K.; Remondino, F.; Yang, M. Y.; Madhuanand, L.; Yogender, Yogender; Alsadik, B.; Weinmann, M.; Jutzi, B.; Qin, R.. - In: ISPRS OPEN JOURNAL OF PHOTOGRAMMETRY AND REMOTE SENSING. - ISSN 2667-3932. - STAMPA. - Volume 13:(2024). [10.1016/j.ophoto.2024.100070]

Availability:
This version is available at: 11583/2989739 since: 2024-06-26T11:58:41Z

Publisher:
Elsevier B.V. on behalf of International Society of Photogrammetry and Remote Sensing (ISPRS)

Published
DOI:10.1016/j.ophoto.2024.100070

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright
Elsevier postprint/Author's Accepted Manuscript

© 2024. This manuscript version is made available under the CC-BY-NC-ND 4.0 license
http://creativecommons.org/licenses/by-nc-nd/4.0/.The final authenticated version is available online at:
http://dx.doi.org/10.1016/j.ophoto.2024.100070

(Article begins on next page)

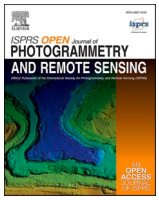
05 January 2025

ISPRS Open Journal of Photogrammetry and Remote Sensing 13 (2024) 100070

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



UseGeo - A UAV-based multi-sensor dataset for geospatial research

F. Nex [a,*], E.K. Stathopoulou [b], F. Remondino [b], M.Y. Yang [a,j], L. Madhuanand [d], Y. Yogender [a,c], B. Alsadik [a], M. Weinmann [e], B. Jutzi [e], R. Qin [f,g,h,i]

[a] ITC Department of Earth Observation Science, Faculty ITC, University of Twente, Enschede, the Netherlands
[b] 3D Optical Metrology (3DOM) Unit, Bruno Kessler Foundation (FBK), Trento, Italy
[c] Inter-university Department of Regional & Urban Studies and Planning (DIST), Politecnico di Torino, Italy
[d] Department of Physical Geography, Faculty of Geosciences, University of Utrecht, Utrecht, Netherlands
[e] Institute of Photogrammetry and Remote Sensing, Karlsruhe Institute of Technology, Karlsruhe, Germany
[f] Geospatial Data Analytics Laboratory, The Ohio State University, Columbus, USA
[g] Department of Civil, Environmental and Geodetic Engineering, The Ohio State University, Columbus, USA
[h] Department of Electrical and Computer Engineering, The Ohio State University, Columbus, USA
[i] Translational Data Analytics Institute, The Ohio State University, Columbus, USA
[j] Visual Computing Group, University of Bath, Bath, UK

ARTICLE INFO

Keywords:
UAV
LiDAR
Monocular depth estimation
Stereo matching
Multi-view stereo
3D reconstruction
Deep learning

ABSTRACT

3D reconstruction is a long-standing research topic in the photogrammetric and computer vision communities; although a plethora of open-source and commercial solutions for 3D reconstruction have been released in the last few years, several open challenges and limitations still exist. Undoubtedly, deep learning algorithms have demonstrated great potential in several remote sensing tasks, including image-based 3D reconstruction. State-of-the-art monocular and stereo algorithms leverage deep learning techniques and achieve increased performance in depth estimation and 3D reconstruction. However, one of the limitations of such methods is that they highly rely on large training sets that are often tedious to obtain; even when available, they typically refer to indoor, close-range scenarios and low-resolution images. Especially while considering UAV (Unmanned Aerial Vehicle) scenarios, such data are not available and domain adaptation is not a trivial challenge. To fill this gap, the UAV-based multi-sensor dataset for geospatial research (UseGeo - https://usegeo.fbk.eu/home) is introduced in this paper. It contains both image and LiDAR data and aims to support relevant research in photogrammetry and computer vision with a useful training set for both stereo and monocular 3D reconstruction algorithms. In this regard, the dataset provides ground truth data for both point clouds and depth maps. In addition, UseGeo can be also a valuable dataset for other tasks such as feature extraction and matching, aerial triangulation, or image and LiDAR co-registration. The paper introduces the UseGeo dataset and validates some state-of-the-art algorithms to assess their usability for both monocular and multi-view 3D reconstruction.

1. Introduction

The generation of complete and accurate 3D representations of scenes using images has been one of the main research topics for the last decades in the photogrammetric and computer vision communities. Various methods with different outputs have been developed toward this scope for diverse applications such as mapping, autonomous navigation, localization, and virtual or augmented reality, among others. Depth estimation and 3D reconstruction algorithms have been largely improved in recent years, as also witnessed by the release of many solutions that are nowadays commonly used by researchers and practitioners in different domains. These solutions were mainly developed following standard Structure from Motion (SfM) and Multi-View Stereo (MVS) pipelines for real-world and large-scale applications and are normally based on conventional approaches using hand-crafted features and user-defined parameters. Despite the impressive results of such solutions, delivering precise, complete, and aesthetically pleasing 3D reconstruction results in multi-view scenarios is still an open challenge

* Corresponding author.
E-mail addresses: f.nex@utwente.nl (F. Nex), estathopoulou@fbk.eu (E.K. Stathopoulou), remondino@fbk.eu (F. Remondino), myy35@bath.ac.uk (M.Y. Yang), l.madhuanand@uu.nl (L. Madhuanand), yogender.yadav@polito.it (Y. Yogender), b.s.a.alsadik@utwente.nl (B. Alsadik), martin.weinmann@kit.edu (M. Weinmann), boris.jutzi@kit.edu (B. Jutzi), qin.324@osu.edu (R. Qin).

https://doi.org/10.1016/j.ophoto.2024.100070
Received 3 September 2023; Received in revised form 3 June 2024; Accepted 17 June 2024
Available online 18 June 2024
2667-3932/© 2024 The Authors. Published by Elsevier B.V. on behalf of International Society of Photogrammetry and Remote Sensing (isprs). This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

for the scientific community. Inevitably, acquisition conditions such as image network geometry, illumination conditions, and sensor quality can severely affect the reconstruction results; yet the efficiency of the implemented algorithm is of utmost importance to ensure high-fidelity outcomes (Seitz et al., 2006; Wenzel et al., 2013; Remondino et al., 2014; Aanæs et al., 2016; Knapitsch et al., 2017).

Deep neural networks are used in several visual recognition tasks such as image classification (Krizhevsky et al., 2012; He et al., 2016), object detection (Girshick et al., 2014; He et al., 2017), and semantic segmentation (Long et al., 2015; Chen et al., 2017; Badrinarayanan et al., 2017) with great success, mainly due to their capability to consider the global semantic context. For depth and disparity estimation, convolutional networks have been exploited in the two-view (Zbontar and LeCun, 2015; Kendall et al., 2017; Guo et al., 2019) and multi-view scenarios (Yao et al., 2018; Im et al., 2019; Xu et al., 2021b). Deep learning has also revitalized the interest in monocular depth estimation algorithms (a.k.a. Single Image Depth Estimation (SIDE) or Multi-view Depth Estimation (MDE)) that have become popular in many indoor and outdoor applications due to their cost effectiveness and flexibility (Eigen et al., 2014; Laina et al., 2016; Godard et al., 2017; Xu and Tao, 2020; Yin et al., 2021). However, despite their undeniable potential, the applicability of these methods in real-world scenarios is still debatable. Among the biggest concerns are the need for a vast amount of training data, the high memory requirements, and the limited generalization and domain adaptation performance. Several benchmarks have already been released by different communities in the last years to promote the development of efficient and reliable algorithms across different applications (Geiger et al., 2012; Mayer et al., 2016; Schöps et al., 2017; Choy et al., 2019; Hermann et al., 2020; Madhuanand et al., 2021; Welponer et al., 2022; Wu et al., XLIII-B2021). The ground truth (GT) in these benchmarks is usually provided by point clouds or surface models obtained using active sensors or a-priori-generated synthetic models of the scene. In photogrammetry and 3D vision, most of the benchmarks focus on satellite and terrestrial datasets (Wu et al., XLIII-B2021). Airborne benchmark data are historically less popular, and although their number is progressively increasing, only a few of these benchmarks have been dedicated to UAV datasets (Nex et al., 2015; Lyu et al., 2020). These UAV datasets are either dedicated to semantic segmentation or aim to assess the image orientation process with ground control points, while the quality of the 3D reconstruction is mainly qualitative. UAV datasets allow ultra-dense 3D reconstructions but their comparison with conventional airborne LiDAR data is normally insufficient to allow a thorough comparison (Nex et al., 2022), given the different point densities of these data.

The aim of the paper is to introduce the UseGeo dataset (https://usegeo.fbk.eu/home) that intends to bridge the aforementioned gaps, providing images and GT point clouds acquired by UAV platforms for the rigorous assessment of 3D reconstruction algorithms, with a specific focus on deep learning approaches. The dataset has been supported by ISPRS (Scientific Initiatives, 2021) and aims to foster research on very high-resolution images, providing a useful training set for both MVS and monocular 3D reconstruction algorithms. Simultaneous acquisition of images and LiDAR was performed in different urban and peri-urban areas. While LiDAR was acquired (primarily but not necessarily) as a reference (GT), the image data sources were used for the training and testing of MVS and monocular 3D reconstruction algorithms. Different typologies of landscapes have been considered in the acquisition to deliver relatively heterogeneous scenes. The data have already been validated using certain meaningful state-of-the-art algorithms to assess the dataset usability for both, monocular and MVS 3D reconstruction tasks. With the availability of image blocks and LiDAR data, UseGeo can be also a useful dataset for additional research tasks such as image triangulation, exploiting novel feature extraction and matching algorithms, and image/LiDAR ~~co-registration and~~ fusion.

The paper is organized as follows. In Section 2 a literature review on MVS and monocular 3D reconstruction algorithms as well as other tasks potentially benefiting from UseGeo is provided. Section 3 reports the used UAV system and the data collection. The pre-processing steps employed to prepare the data are summarized in Section 4, while some preliminary tests on state-of-the-art algorithms are reported in Section 5. The data organization for their efficient download is finally reported in Section 6.

2. State–of–the-art

In the following sections, we report the tasks where UseGeo data could be mainly used to develop and evaluate algorithms and methodologies, in particular: multi-view depth estimation and 3D reconstruction and monocular depth estimation. The described dataset could be additionally used to assess other tasks such as image orientation, feature extraction and matching or even automated registration of images and LiDAR data.

2.1. Multi-view depth estimation and 3D reconstruction

The SfM process typically relies on matching local features and yields an abstract representation of the scene consisting of few but high-fidelity 3D points along with the camera orientations. MVS algorithms, on the other hand, aim to generate a complete and dense 3D representation of the scene, either as a point cloud or a triangulated mesh (Zhou et al., 2020; Wang et al., 2021; Stathopoulou and Remondino, 2023). Thus, the depth calculation of every pixel is attempted to establish robust pixel correspondences. In the standard two-view scenario, epipolar geometry constraints simplify the correspondence search by restricting the search space along the epipolar line. Several methods have been developed for solving this correspondence search problem, either local (Scharstein, 1994; Hosni et al., 2011; Bleyer et al., 2011), global (Faugeras and Keriven, 1998; Strecha et al., 2004), or hybrid semi-global (Hirschmüller, 2008) methods. Nonetheless, MVS scenarios, considering multiple images of the same scene, mostly refer to non-rectified images. MVS relies on the same principles for correspondence search but has a higher complexity than its two-view equivalent due to the ray redundancy resulting from the multiple observations and the arbitrary viewpoint variations. Many specially designed algorithms have been developed for efficient MVS reconstruction in recent years, having achieved impressive results (Strecha et al., 2006; Galliani et al., 2015; Schönberger et al., 2016; Schönberger and Frahm, 2016b; Xu and Tao, 2019). A popular categorization for MVS approaches is based on their reconstruction algorithms (Seitz et al., 2006), which are voxel-based methods, surface evolution-based methods, feature point growing-based methods, and depth map-based methods. Depth map fusion algorithms perform per-view depth estimation and subsequent depth fusion (Strecha et al., 2006; Merrell et al., 2007; Gallup et al., 2007; Furukawa et al., 2010; Galliani et al., 2015). They have been widely used in large-scale, high-resolution applications with demanding accuracy requirements due to their overall efficiency and scalability. Regarding learning-based methods for depth estimation under MVS scenarios, early approaches in the field were based on volumetric scene representations and learned voxel occupancy (Ji et al., 2017; Kar et al., 2017; Paschalidou et al., 2018), but their applicability was limited to low-resolution scenarios due to high computational complexity. More recent methods adopted plane-sweep volumes to enable better scalability and achieve impressive time results (Huang et al., 2018; Yao et al., 2018). To confront, up to some extent, the high memory needs of 3D cost volumes, RNN (Recurrent Neural Networks) architectures have also been implemented (Yao et al., 2019; Xu et al., 2021) along with coarse-to-fine schemes (Yang et al., 2020). Nevertheless, learning-based MVS methods have limited applicability in high-resolution datasets due to the use of 3D convolutions for cost volume representation and they are typically evaluated over low- or medium-resolution datasets followed by refinement and postprocessing steps. Moreover, such methods, in order to generalize appropriately, need a large amount of GT depth maps for training,

therefore commonly provided by synthetic datasets, e.g., Yao et al. (2020). However, methods trained only on synthetic data inevitably suffer from domain differences with real-world scenarios. To relax the requirement of GT depth maps for training, unsupervised methods have also been exploited, considering the nearby views (Khot et al., 2019; Dai et al., 2019) and additional semantic cues (Xu et al., 2021). However, they still show insufficient performance and scalability for large-scale scenarios.

2.2. Monocular depth estimation

Monocular Depth Estimation (MDE) refers to the process of recovering distances between objects in the 3D space along with the camera parameters given only one image. Since limited information about the scene structure can be directly extracted from a single image (ill-posed problem), prior cues regarding the captured scene should be provided for efficient 3D reconstruction. Early methods for monocular depth estimation relied on handcrafted features and used such complementary cues to recover the depth by formulating an MRF (Markov Random Field) (Saxena et al., 2008). Using deep learning methods, monocular depth prediction refers to the single image inference during test time and is typically formulated either as a regression or a classification problem. Eigen et al. (2014), in a seminal work, proposed a coarse-to-fine scheme and a scale-invariant loss function. As follow-up work, Eigen and Fergus (2015) also predicted surface normals and semantic maps in a similar framework. Various methods have been proposed in the recent literature that consider monocular depth estimation as a supervised (Laina et al., 2016; Xu et al., 2018; Fu et al., 2018; Hu et al., 2019) or a self-supervised problem (Garg et al., 2016; Godard et al., 2017; Tosi et al., 2019; Zhang et al., 2023). In supervised methods, GT is often obtained by sparse depth maps generated using LiDAR point clouds, since rich GT depth annotations are costly to obtain for every pixel. For self-supervised methods, on the other hand, binocular cues such as left-right consistency are used to circumvent the need for GT data, while a research direction focuses on training from monocular video, considering motion (Zhou et al., 2017; Teed and Deng, 2019). Loss functions are formed either based on pixel-wise photometric loss, either L1 or L2, (Garg et al., 2016), or by combining more sophisticated cues such as Structural Similarity Index (SSIM) (Godard et al., 2017; Watson et al., 2019) to measure the similarity of two image patches. Depending on the available training data, the scene depth can be estimated as ordinal, i.e., relative (Fu et al., 2018) or Euclidean (Eigen et al., 2014; Yin et al., 2019). Despite achieving impressive results in depth map inference in common benchmarks (Silberman et al., 2012; Geiger et al., 2012), the respective 3D reconstructions suffer from apparent distortions and artefacts. Some recent works attempt to incorporate 3D geometry cues in monocular depth estimation; for instance, plane priors can be considered based on the assumption that urban scenes are commonly composed of planar structures (Lee et al., 2019). Yin et al. (2019) formulated a joint loss function using virtual normals to explicitly consider the 3D structure and enforce high-order geometric consistency between surface patches in a large range. Further extensions of this work refer to affine-invariant depth formulation (Yin et al., 2020) and add an extra training module for scene 3D reconstruction (Yin et al., 2021). These state-of-the-art methods, although promising, still suffer from limited domain adaptation capability. In a recent method, the potential of monocular depth estimation for 3D reconstruction in photogrammetric scenarios showed limited generalization ability (Welponer et al., 2022; Zhang et al., 2023).

2.3. Other tasks: feature extraction and image-LiDAR registration

The extraction of accurate and reliable tie points among images is the first and fundamental step for the accurate recovery of camera parameters and the 3D reconstruction of the scene. In the last three decades, an incredible number of algorithms have been developed (Gonzalez-Aguilera et al., 2020) to detect homologous points in an automated way. Traditional algorithms were hand-crafted descriptors and detectors, while an increasing number of deep learning approaches are implementing solutions where convolutional networks are adopted to learn directly from the data which features are more effective in the extraction of homologous regions across images (Remondino et al., 2021). Depending on the implementation, different approaches can be categorized: learning-based detectors (Savinov et al., 2017) to only extract features, learning-based detectors (Ebel et al., 2019) to describe the regions around these features, and descript-detect solutions (Dusmanu et al., 2019) or end-to-end learning (Luo et al., 2020) where both detection and description are performed simultaneously. Despite their great potential, these solutions still have significant limits in handling geometric, radiometric and scale changes and, last but not least, deliver different exterior orientation performances depending on the way they were designed (Revaud et al., 2019). In addition, the great majority of these methods have been trained and tested on terrestrial and close-range images, while airborne data are largely overlooked (Remondino et al., 2022).

The integration of photogrammetric blocks or single images with LiDAR point clouds can be useful in different applications ranging from 3D building modeling to object detection and segmentation. The acquisition of images and point clouds can be performed on the same platform (contemporarily) or in separate acquisitions. In both cases, the georeferencing and the synchronization provided by onboard instruments are insufficient to guarantee the accurate alignment of these data without any post-processing step, especially in the case of ultra-high-resolution data acquired by UAVs (Nex et al., 2022). In this regard, several registration methods have been implemented in the last decade to register photogrammetric blocks and LiDAR data. (Abayowa et al., 2015) used the Iterative Closest Point (ICP) algorithm to align different point clouds generated using LiDAR and photogrammetry, while more recently other methods proposed a similar solution using deep learning approaches (Zhang et al., 2020). Similarly, (Yang and Chen, 2015) minimized the discrepancies between point clouds using building outlines and applying a rigid transformation (Xu et al., 2023) while (Toschi et al., 2021) the alignment was performed using points extracted from both datasets: this approach showed good results with already georeferenced point clouds but had limitations using different point cloud densities or acquisitions performed using different platforms. With multi-sensor systems, other approaches combine the LiDAR strip adjustment and the photogrammetric Bundle Block Adjustment to exploit the common GNSS/INS (Global Navigation Satellite System/-Inertial Navigation System) trajectory recorded during the data acquisition. In that regard, solutions proposed by various authors (Glira et al., 2019; Haala et al., 2020, 2022; Zhou et al., 2021) perform tighter integrations of GNSS/IMU, images and LiDAR data by extracting common features from LiDAR and photogrammetric point clouds and using their matches to improve the bundle adjustment, correcting the trajectories of the sensors and improving the alignment.

The registration between a single image and a point cloud has been another relevant topic in the scientific community (Kaminski et al., 2009) but it has been recently boosted by the introduction of deep learning methods. (Li and Lee, 2021) estimated the rigid transformation between coordinate frames converting this into a CNN classification problem to learn common feature descriptors to establish correspondences while (Yan et al., 2022) incrementally aligns the image to the point clouds using a reinforcement learning approach that learns how to improve the pose movements towards the final solution. (Rotstein et al., 2022) aligns coloured point clouds and RGB images minimizing the photometric difference between the colors of the point cloud and those of the corresponding pixels in the image. All these approaches mainly focus on terrestrial applications and data, while their use on airborne (and UAV) images do not exist yet.

3. Data collection

The data were acquired with a RIEGL miniVUX-3UAV scanner and a SONY ILCE-7RM3 camera (Fig. 1) that installs an APX-20 IMU system onboard and guarantees high accuracy in the georeferencing. Using this setup, a total area of 1100 m × 650 m was acquired during a campaign over the Italian territory in April 2021. The average height above ground was 80m with a GSD of approximately 2 cm and the laser point cloud counting 51.39 points/m^2. In some areas, higher point density was achieved. Each acquisition was performed on average with 80% and 60% forward and side image overlap respectively. This overlap guarantees a minimum of 8 images on each object point, with less than 2 cm GSD and an average 140 m × 95 m footprint. Trajectories were initially corrected using riPRECISION and its pre-processing steps, allowing to adjust the trajectories and merge the overlapping strip acquisitions. In total, three flights were performed (Fig. 2), acquiring a total of 829 images (see Table 1).

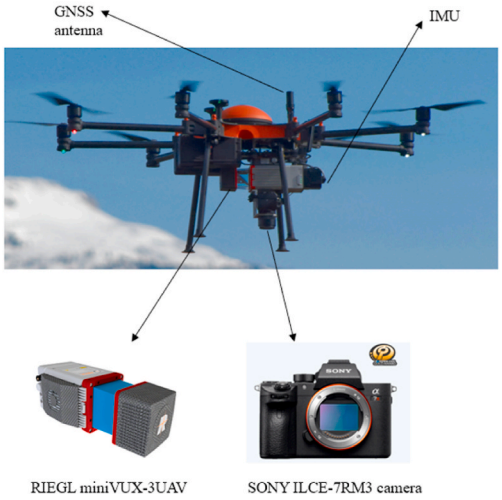
4. Pre-processing of the data

The acquired UseGeo data were processed using different steps in order to guarantee the quality of the outputs and enable their usability by the community. In the following sections, a more detailed description of each step is provided.

4.1. Data alignment

The initial dataset is composed of data acquired by four sensors: LiDAR, optical camera, GNSS, and IMU. The LiDAR strips (.rdbx) are given in the Scanner Coordinate System (SCOS), the initial level arm of the camera images is given with respect to the GNSS/IMU sensors, and the initial GNSS/IMU trajectories and attitudes of the drone are provided in a cartographic reference system. The hybrid adjustment (Pfeifer et al., 2014) approach was then used as an efficient approach to align camera and LiDAR datasets without using any additional ground truth inputs in the form of Ground Control Points (GCPs) or Control Point Clouds (CPCs). The concept of the hybrid adjustment approach is to simultaneously optimize the orientation of LiDAR and the camera by minimizing the discrepancies between the produced point clouds. The camera images are pre-processed to obtain the exterior orientations (EO), image point observations, and tie points needed for the hybrid adjustment. In the hybrid adjustment, correspondences were established and selected between image pairs (IMG-IMG), overlapping LiDAR strips (STR-STR), image tie points and LiDAR strips (IMG-STR) with a modified Iterative Closest Point (ICP) algorithm. The subsets of the points are selected from the correspondences with a uniform sampling technique to make the adjustment process computationally efficient. Uniform sampling aims to select the points from both datasets in the object space as consistently as possible (Glira et al., 2015a) and to ensure that the uniform distribution of points in the correspondences and equal-area regions are weighted equally within the hybrid adjustment for the implementation of uniform sampling of the points. The regions of overlap were divided into a voxel structure and the nearest point to each voxel centre was selected. The edge length of each voxel can be treated as the mean sampling distance along each coordinate direction (Glira et al., 2015a). Fig. 3 illustrates the following steps for aligning the datasets using hybrid adjustment. The readers may refer to (Glira et al., 2019) for more technical details.

The main iteration loop in the hybrid adjustment starts with the direct georeferencing of the LiDAR strips with the initial parameters in the first loop and then continues the refinement of the estimated parameters from the hybrid adjustment in the subsequent loops. During the adjustment procedure, the potential correspondences are matched, i.e., the nearest neighbour of a query point in the overlapping point cloud. The false correspondences are rejected and removed in the subsequent step based on threshold criteria (Glira et al., 2015b) of the roughness, the angle between the normal vectors of corresponding points, and the distance between the corresponding points (Glira et al., 2015a, 2015b). After the outlier rejection step, the correspondences are weighted based on their surface roughness and angle between respective surface normals. It is worth mentioning that the correspondences are recalculated in each iteration of hybrid adjustment. After a given number of iterations are completed, the LiDAR strips are georeferenced with the adjusted parameters in the final iteration loop of the hybrid adjustment. As a final product of the hybrid adjustment, adjusted LiDAR strips, adjusted image orientations, camera calibration and undistorted images



(a)

MiniVUX-3

Scan speed: up to 100 scans/sec

Measurement rate: up to 200,000 meas./sec

Field-of-view: 360°

Weight: 1.55 kg

Target echoes / laser shot: up to 5

SONY ILCE – 7RM3

Sensor: full frame (35 mm)

Sensor size: 7952 x 5304 pixels (35.9 x 24 mm)

Focal length: 28-70 mm

Weight: 657 g

Size:126.9 mm x 95.6 mm x 73.7 mm

(b)

Fig. 1. The used drone (a) and onboard sensors technical specifications (b).

Fig. 2. Photogrammetric point clouds generated on the test areas.

Table 1
Collected datasets and the corresponding number of images.

	Strips	# images	Forward overlap	Side overlap	GSD [cm]
Dataset 1	8	224	80	60	1.7
Dataset 2	8	328	80	60	1.8
Dataset 3	8	277	80	60	1.9

The high overlap between images has guaranteed the generation of high-quality and complete point clouds that have been used to verify the alignment between photogrammetric and LiDAR data (see Section 4.1).

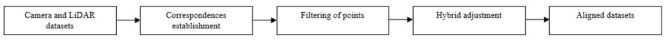


Fig. 3. LiDAR and camera data alignment process.

as well as the adjusted trajectory of the UAS are obtained. The data that can be downloaded in UseGeo refers to this output.

For results evaluation and analysis of the achieved results, the camera images with their adjusted orientations were processed in Pix4DMapper software to generate a dense point cloud (Yadav et al., 2023). The primary quality analysis was carried out in CloudCompare software with the computation of the mean cloud-to-cloud (C2C) distances between LiDAR and camera point clouds, as shown in Fig. 4 for Area 1.

The regions near the block borders usually had higher residuals, which can be attributed to the regulatory behavior of the hybrid adjustment process implemented on the dataset. However, it is quite evident that the initial alignment error (above 1 m in all the three test areas) has been significantly reduced thanks to the hybrid adjustment (Fig. 4). The C2C distances were very similar in all three test areas,

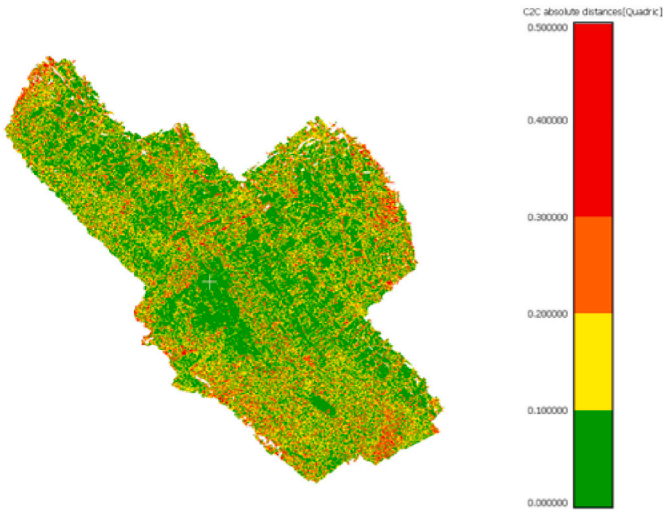


Fig. 4. Example of mean cloud-to-cloud distances between LiDAR and camera point clouds for area 1, after the hybrid adjustment.

guaranteeing residuals equivalent within 2–3 image GSDs (and within the distance between points in the LiDAR data) in all the performed adjustments, as reported in Table 2.

4.2. Depth map generation

The corresponding projection matrix P for each image is obtained from the adjustment described above, encapsulating the extrinsic and intrinsic parameters. Given this information along with the GT point cloud acquired with the LiDAR, we generate a GT depth map for each image by back-projecting the 3D points into the image plane. The undistorted images given in output by the adjustment were used in this step. In particular, the Z-buffer algorithm (Habib et al., 2007) was adopted for the back-projection of the depth values to account for occlusions in the scene and prevent double-mapping problems. As well known, Z-buffer has problems handling occlusions, especially when range and image acquisitions are performed from different positions. This problem is even more frequent in correspondence of depth discontinuities (e.g. buildings) and when LiDAR data are relatively sparse compared to the image resolution. In the UseGeo dataset, this issue was mitigated by acquiring data with the same platform, by the high density of the LiDAR point cloud and by down-sampling the depth images (to reduce the size of the depth image too). In addition, the residual errors were removed by applying a 3x3 matrix filtering to the depth values lower than the median (<2.5 m) values. These points as well as other residual empty pixels in the scene were filled using a linear interpolation. Because of the limited extension of the areas to fill, more complex interpolators showed comparable results. An example of this process is shown in Fig. 5.

Please note that the delivered depth files report the depth as Euclidean distances between the projection centre of the camera and the object space (*depth* in Fig. 6, a). In many algorithms (such as self-supervised approaches) the depth learnt by the network is the distance between the camera and the object space in Z direction (like an orthographic projection), considering the Z axis parallel to the optical axis of the camera (*Z-depth* in Fig. 6, a). To consider these two different depth definitions, a correction file (8-bit.tif files) has been computed for each dataset. The values (Digital Number, DN) reported in this file range between 0 and 255 (Fig. 6, b) and vary according to each pixel distance from the camera's principal point. The conversion from the Euclidean to the Z-depth distances can be achieved by multiplying their normalized value (i.e. DN/255) by the Euclidean depth values. In the considered dataset, this ratio varied from 1 (central part of the image to about 0.75 on the corner of the image.

Table 2
Mean values of cloud-to-cloud distances between LiDAR and photogrammetric point clouds after hybrid adjustment.

	Mean C2C distances (cm)
Dataset 1	8.8
Dataset 2	8.5
Dataset 3	6.7

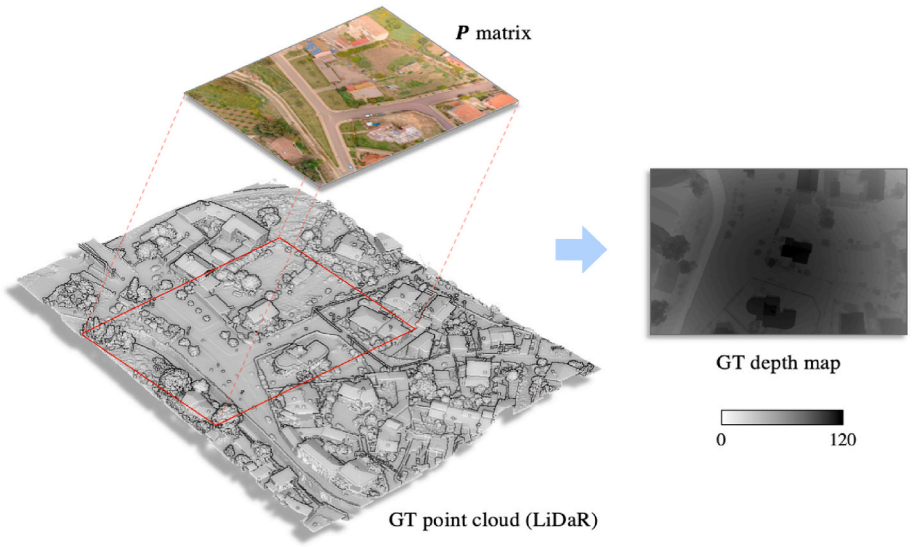


Fig. 5. For each image of known orientation parameters, a GT depth map is generated by projecting the LiDAR 3D points to the image plane.

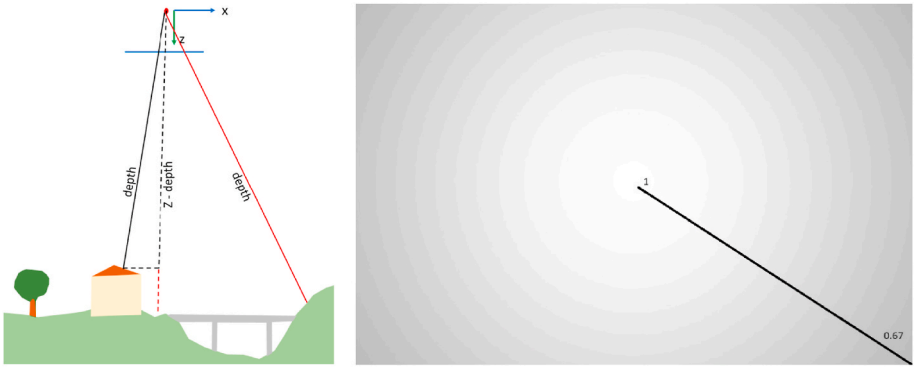


Fig. 6. (a) Scheme showing the difference between (Euclidean) depths and Z-depths: depth difference increases with the distance from the principal point of the image (see black and red line); (b) example of conversion file from Euclidean to Z-depth maps: brightest colors refer to values close to 1 (no difference between Euclidean and Z-depths) while grayer colors defines lower values (bigger differences between Euclidean and Z-depths) on the edges of the image (value 0.67). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

5. Tests and results

The delivered data have been tested using state-of-the-art algorithms to provide examples of the achievable results using the presented benchmark. In particular, the results achieved by SIDE algorithms and stereo-matching algorithms are shown in the following. As already mentioned, a comprehensive evaluation of the state-of-the-art methods for single-depth estimation or stereo reconstruction is out of the scope of this paper. The tests reported below do not follow the same procedure and, therefore, the achieved results cannot be directly compared. A more detailed evaluation of state-of-the-art methods is reported in (Nex et al., 2023) and in (Hermann et al., 2024).

This paper aims to demonstrate the usability of the released data for such purposes, with specific regard to deep learning methods that require larger datasets for training. The results reported in the following indicate some of the possible tasks where this benchmark will be useful for the scientific community.

5.1. Performance assessment

The metrics reported in Eqs. (1)–(4) were adopted to assess the tested algorithms. These metrics compare the estimated depth (d') generated by the algorithms and their corresponding ground truth depth (d), averaging these values on the number of pixels N of each depth map. In particular, the Absolute Relative difference (Abs Rel) (often called L1-rel) given in equation (1), Squared Relative difference (Sq Rel), equation (2), Root Mean Square Error (RMSE), equation (3), were used in this assessment as also described in Godard et al. (2019) and Hermann et al. (2020).

$$\text{Abs Rel} = \frac{1}{N}\sum_{i=1}^{N}\frac{|d(x_i) - d'(x_i)|}{d(x_i)} \tag{1}$$

$$\text{Sq Rel} = \frac{1}{N}\sum_{i=1}^{N}\frac{|d(x_i) - d'(x_i)|^2}{d(x_i)} \tag{2}$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(d(x_i) - d'(x_i))^2} \tag{3}$$

$$\text{Accuracy } (\delta\theta) = \frac{1}{N}\sum_{i=1}^{N} max\left(\frac{d_i}{d_i'}, \frac{d_i'}{d_i}\right) < \theta \tag{4}$$

The accuracy δ (equation (4)) reports the percentage of pixels that do not exceed a certain difference from the corresponding pixel value in the reference depth map (Godard et al., 2019). This difference is measured considering the maximum among the ratios of the computed depth and its corresponding ground truth and their inverse value. The thresholds θ are 1.25, 1.15 and 1.05 as also proposed in the KITTI standard benchmark (Garg et al., 2016).

5.2. Monocular depth estimation (MDE)

Two different algorithms have been used in this assessment. The first algorithm refers to (Madhuanand et al., 2021), a self-supervised approach not requiring ground truth depth maps for training, instead, using sequences of frames to learn both depths and pose information through two different networks. The predicted depth and pose are jointly exploited to reconstruct one image from the viewpoint of another image of the same dataset, adopting a contrastive loss term to improve image quality generation. More in detail, the network combines 2D encoders and 3D decoders for extracting information from adjacent images. In (Madhuanand et al., 2021), the model was trained using oblique images acquired by UAV. The first tests performed using this previously trained model on nadir images did not provide acceptable results. It was then decided to retrain the model using a combination of images from UseGeo, the Hessigheim3D (Kölle et al., 2021) and Zeche Zollern (Nex et al., 2015) datasets.

The images from Hessigheim3D and Zeche Zollern datasets are of size 6132 × 8176 and 1989 × 1320, respectively, with an overlap of 80% with consecutive images. While both datasets comprise features like rooftops, vegetation, roads, and barren land, the Zeche Zollern dataset appears to be more of a densely populated urban setting than the Hessigheim3D dataset. A total number of 1036 images was used for training, 136 images for validation, and 88 images for testing; the testing was performed using only images from the UseGeo dataset.

The architecture was implemented using the PyTorch framework (Paschalidou et al., 2018) and trained using resized input images of resolution 640 × 352 pixels. The learning rate was set to 10^{-5} and the Adam optimizer (Kigma et al., 2014) was used for optimization. The training lasted 40 epochs with a batch size of 12. The weights between different loss terms are used as in Madhuanand et al. (2021). We used a single Nvidia Titan Xp GPU with 16 GB memory, with a total computation time of ca. 11 h. To assess the performance of the method, different pixel-wise metrics are calculated between the predicted and reference depth according to the same rationale adopted by Madhuanand et al. (2021). The achieved results are summarized in Table 3, while visual examples of the results can be seen in Fig. 7.

From the inferred depth maps it can be observed that the edges of buildings, roads, and other structures are obtained with a higher quality in comparison to the trees and objects that are closer to ground level. Overall, the reconstruction is still very blurry, ignoring finer details. Another possible limitation of the achieved results is given by the different GSD of the subsampled image (both resized at the same scale but starting from different image sizes and GSDs) that could have reduced the quality of the 3D depth estimation.

The second algorithm that was used to test the benchmark refers to (Hermann et al., 2020) and also represents a self-supervised learning approach for single image depth estimation. This approach is trained by executing four consecutive steps in each training iteration: 1) An encoder-decoder network is used for monocular depth estimation to predict a depth map for the reference image. 2) A pose estimation network is used to estimate the relative orientation between the reference image and two adjacent images in the image sequence. 3) A Spatial Transformer Network together with bilinear sampling is used for synthetic view generation from the two adjacent images using the estimated depth map and relative camera poses. 4) An image reconstruction error is calculated by comparing the synthetic images and the original image. This error serves as the training loss and comprises a photometric loss to enforce a high-quality image reconstruction as well as an edge-aware smoothness loss to enforce smooth depth maps. Once the approach has been trained, inference can be performed on only one image by using only the network for monocular depth estimation to predict the corresponding depth map.

Table 3
Quantitative results achieved by the MDE reconstruction approach presented by (Madhuanand et al., 2021). For more information on the metrics used, please refer to (Madhuanand et al., 2021; Garg et al., 2016).

Method	Abs Rel	Sq Rel [m[	RMSE [m]	$\delta_{1.25}$ ↑	$\delta_{1.15}$ ↑	$\delta_{1.05}$ ↑
Madhuanand et al. (2021)	0.049	0.377	5.967	0.999	0.968	0.579

The first tests of this approach on the UseGeo dataset were performed using 510 images for training and 192 images for testing, whereby image selection was focused on reducing redundancy and ensuring non-overlapping training and test areas. All images were rescaled to a size of 768 × 448 pixels. For details on performance evaluation and implementation, we refer to (Hermann et al., 2023). For the sake of reproducibility and comparison, applied evaluation procedures are released in (Hermann et al., 2023). All experiments are conducted on 2 T V100 GPUs. The total processing time is about 30 h for training (also due to the use of data augmentation in order to increase the given amount of training data) and 14 ms per image for testing using a single GPU. The achieved results are summarized in Table 4, and visualizations of predicted depth maps as well as the corresponding reference image and the ground truth depth maps are provided in Fig. 8.

The predicted depth maps reveal that the self-supervised MDE approach of Hermann et al. (2020, 2023) is capable of learning monocular depth estimation from aerial imagery. The scene geometry and objects are predicted correctly, yet the prediction of fine structures (e.g., given for vegetation and sharp edges) remains challenging.

In both experiments, the only use of nadir images for the training of self-supervised approaches could have represented an additional challenge for the successful inference of the network. This is a specific problem that will deserve more attention from our research community in the upcoming years.

5.3. Stereo/multi-view stereo matching

A hybrid multi-view stereo and deep learning method fully exploring the potential of this dataset is assessed. It uses a combination of traditional multi-view stereo (MVS) paradigm based on semi-global matching (SGM) (Hirschmüller, 2008) followed by a refinement network using a few images from UseGeo as the training set. For this MVS algorithm, a connectivity matrix was first established computing the neighbourhood views for each image and using heuristics such as the distance of the perspective centre, intersection angle, overlap, as well as the number of feature points (sparse points). Then a pair-wise SGM algorithm, which is implemented in the software Multi-view Stereo Processor (Qin,), was performed thanks to a hierarchical matching strategy to facilitate large frame images. The MVS process generates, for each view, a number of pairwise depth maps, and these depth maps are then fused with a median filter to generate an initial depth for each view.

Given that UseGeo provides the reference depth data for each view, it can provide a valid means to improve traditional MVS algorithms through depth refinement. As a second step, a U-Net model (Ronneberger et al., 2015) was directly used to regress a 4-channel input, consisting of the three RGB image channels and the initial depth from the MVS algorithm as the 4th channel, to the refined depth. Given that U-Net is a relatively lightweight network, only 10 images from the UseGeo dataset were used for training.

During the training, standard data augmentation such as random rotation and flip was used, and additionally, the images were cropped

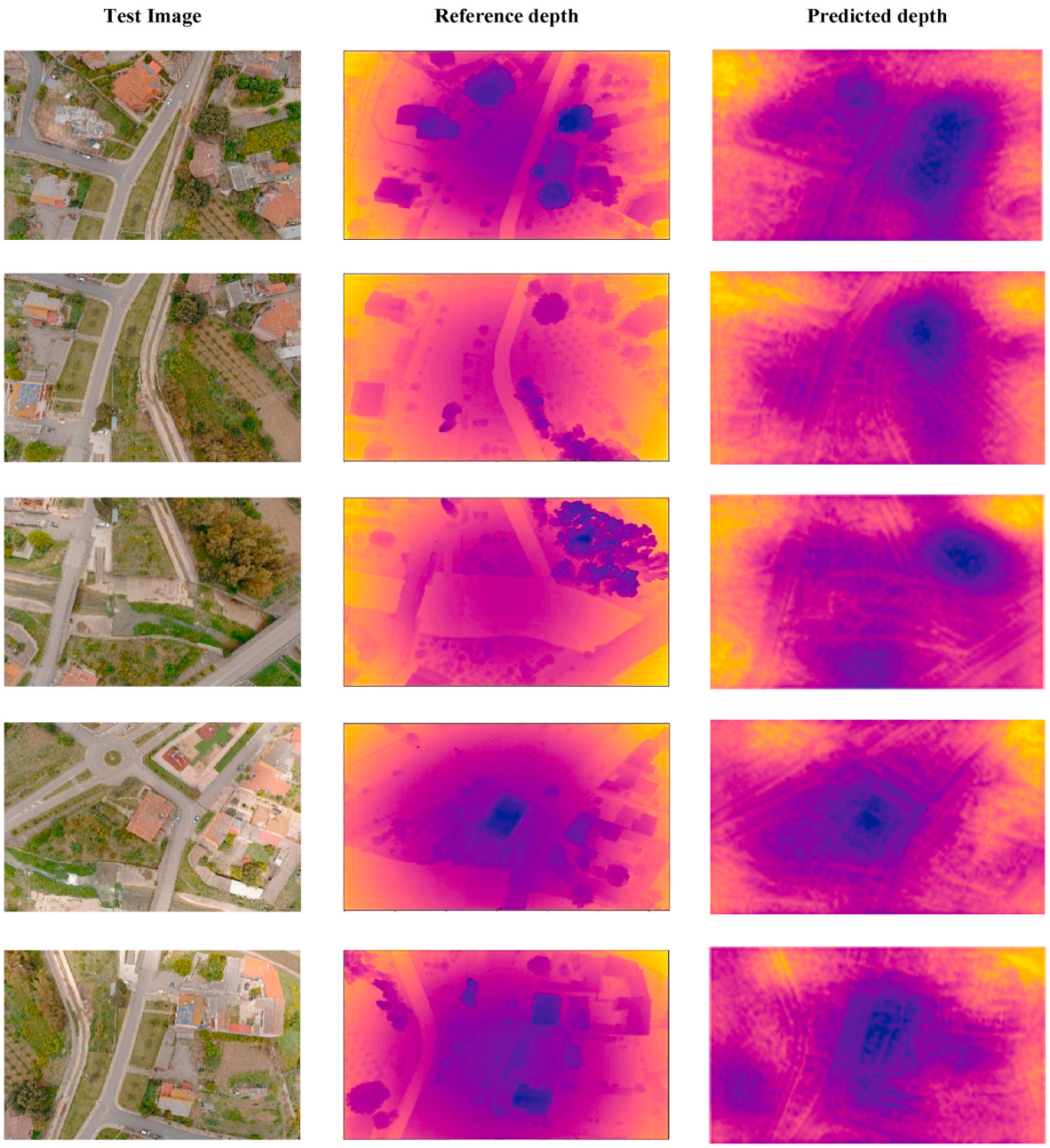


Fig. 7. Examples of delivered results: (left) input image, (center) ground truth provided by the UseGeo dataset and (right) inferred depth from the single image. Note that violet refers to closer objects, while yellow to more distant ones. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Table 4
Quantitative results achieved by the MDE approach presented by Hermann et al. (2020, 2023). L1-rel refers to the relative L1-norm.

Method	Abs Rel	$\delta_{1.10} \uparrow$	$\delta_{1.05} \uparrow$	$\delta_{1.01} \uparrow$
Hermann et al. (2020, 2023)	0.0614	0.9399	0.7436	0.1994

into 256 × 256 patches. In order to train the model with image patches, the depth values were converted to depth with respect to the camera focal plane (i.e. Z-depths). The input was normalized using a global standard deviation and mean values for each channel. A minor adaptation was used to enhance the edges of the depth by giving a higher weight to the depth boundaries when training, where the boundaries were detected using the Canny operator (Canny, 1986). Both L1-loss and gradient loss were used, optimized by Adam optimizer with the initial learning rate 0.01, and 10^{-5} wt decay. Given the lightweight of the refinement network, it can be trained by a relatively less powerful GPU (GTX 1070) for about 10 h, and with an inference time of less than half a second for each image.

Indicative results for 10 randomly selected, overlapping images are shown in Table 5 below. It can be seen that the predicted depth images have revealed better image depth boundaries owing to the additional step of depth refinement benefited from UseGeo. The reconstruction looks complete in all the scene elements and the residuals are definitely lower than the single-image cases, as expected (Fig. 9).

6. Conclusions and released datasets

The UseGeo data are publicly available to the research community to

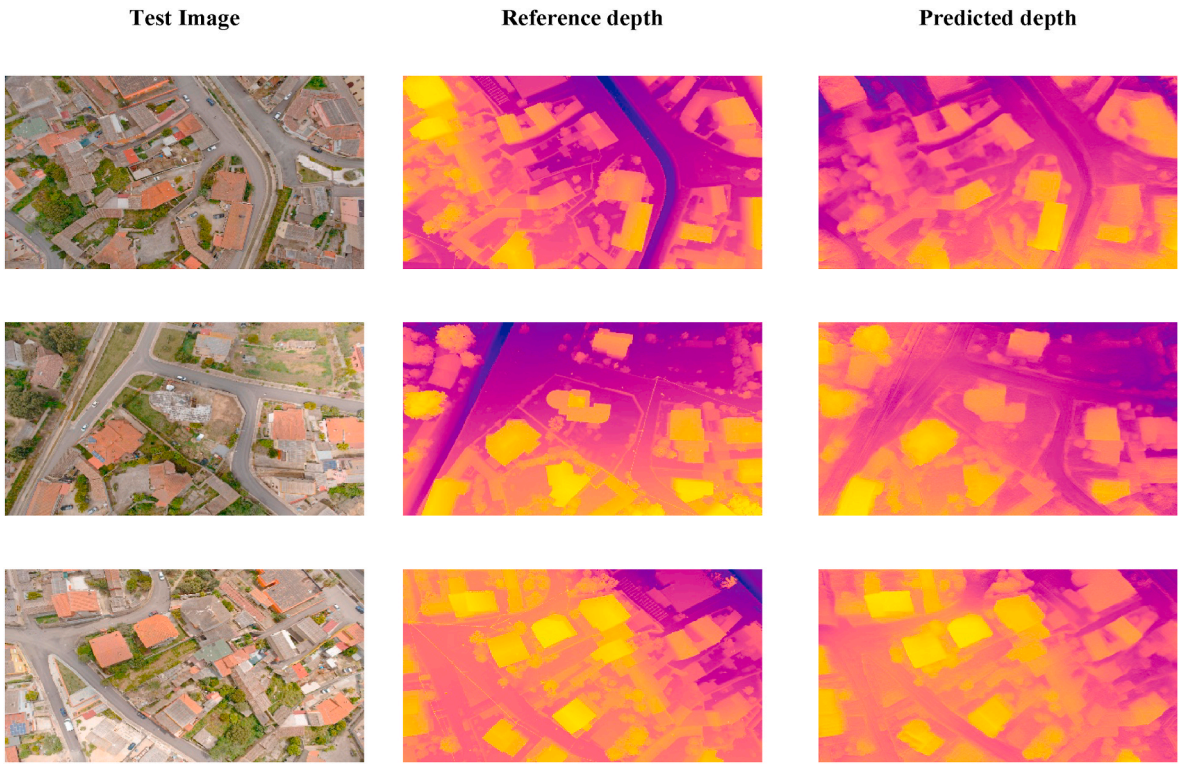


Fig. 8. Visual results achieved by the MDE approach presented by Hermann et al. (2020, 2023): (left) input image, (center) ground truth provided by the UseGeo dataset and (right) inferred monocular depths. Note that yellow refers to closer objects, while violet to more distant ones. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Table 5
Quantitative results achieved by the multi-view stereo approach presented by (Qin,)

Stereo Matching	Abs Rel	Sq Rel [m]	RMSE [m]	$\delta_{1.25}\uparrow$	$\delta_{1.15}\uparrow$	$\delta_{1.05}\uparrow$
Hybrid SGM + DL	0.007	0.025	1.566	0.999	0.998	0.979

promote the development of learning-based approaches for depth estimation on images collected by UAV platforms. It was decided to release all the acquired and processed data to enable different uses and research by the scientific community. For this purpose, the aligned photogrammetric and LiDAR point clouds as well as the full-resolution images block coupled with the image orientation parameters are provided: different formats for the interior and exterior orientation parameters are delivered to ease their use. For each image, a downsampled image (¼ of the original size) and the respective depth map are delivered too. Standard formats are used to distribute the data: specifically, .las files for point clouds, 8-bit.tiff for images and 16-bit.tiff images for depth maps. In addition, the GNSS/IMU trajectories as well as the raw LiDAR files (.rdbx) are released to open new research uses for this dataset. Due to the size of the files, they have been organized into several subfolders to reduce the amount of data to download for each use case. The instructions on how to download the data are available on the UseGeo website: https://usegeo.fbk.eu/home.

6.1. Benchmark's evaluation

UseGeo data can be useful for different research topics, ranging from image orientation to single-image and multi-view stereo 3D reconstruction (Hermann et al., 2024) also considering new solutions such as NeRF (Neural Radiance Field) approaches (Remondino et al., 2023). Furthermore, the additional raw data could make the use of our datasets for additional tasks easier as discussed above. For these reasons, it was decided to not produce strict protocols for results delivery but to leave it open to each research group that wants to publish them. In the UseGeo's website a page and a GitHub folder are dedicated to this. Each research group can freely describe (max 500 words) the used procedure, report the used metrics and (when available) add a few images showing the achieved results.

6.2. Recommendations and future developments of the benchmark

The state-of-the-art algorithms used to assess the suitability of UseGeo have shown that this dataset is a valuable asset to support the improvement of these methods, giving new indications of their performances. The residuals of the performed tests show that there is still room for improvement and this benchmark can support the research to improve these solutions in the upcoming years.

UseGeo also delivers registered images and point clouds that can be freely adopted for a number of possible new applications that go beyond the initial scopes of this scientific initiative. The training and testing of feature extraction and matching algorithms can be developed considering the known orientation of the images and the corresponding coordinates of these features in the object space; analogously, the availability of accurately registered data allows testing automated registration algorithms of images and point clouds. In the following months, the semantic information will be added to the available dataset to increase the number of tasks where UseGeo could be potentially used. In the literature, there are not many benchmarks designed to support different tasks using the same data, especially if we consider ultra-high-resolution images such as UAV data. For this reason, besides the only segmentation task, this addition to UseGeo could also ease the development of multi-task approaches.

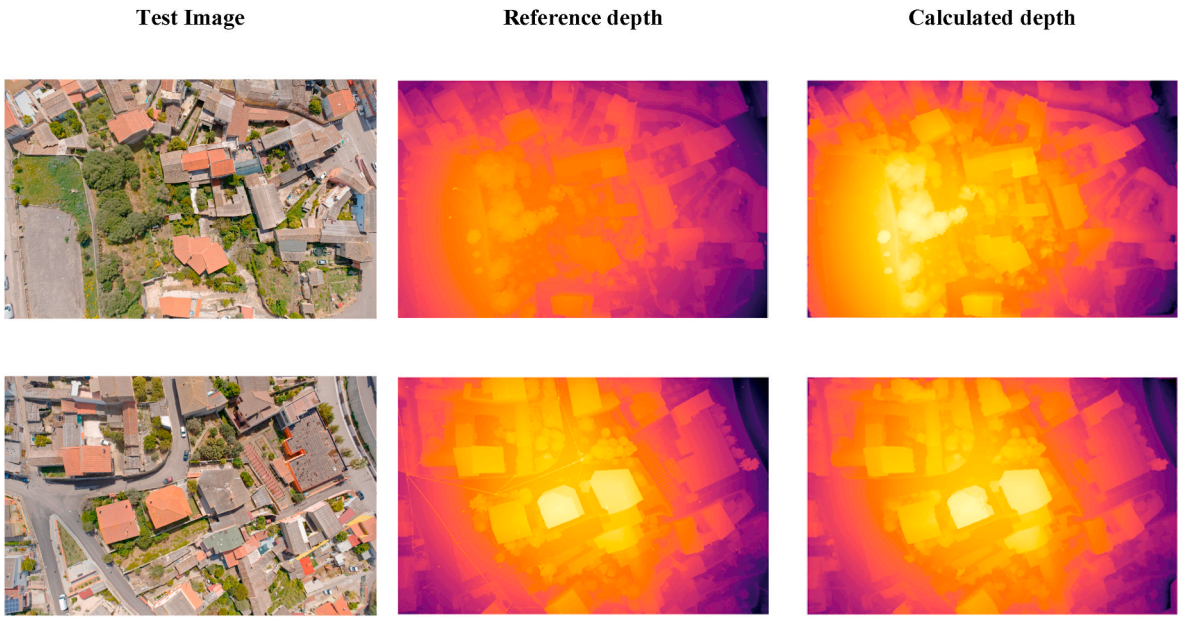


Fig. 9. Qualitative results achieved by the MVS approach presented (Qin,): (left) input image, (center) ground truth provided by the UseGeo benchmark and (right) inferred depths. Note that yellow refers to closer objects, while violet to more distant ones. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Funding

This paper was supported by ISPRS and funded by the ISPRS scientific initiative 2021.

CRediT authorship contribution statement

F. Nex: Writing – original draft, Supervision, Project administration, Methodology, Funding acquisition, Formal analysis, Data curation, Conceptualization. **F. Remondino:** Writing – review & editing, Resources, Project administration, Methodology, Funding acquisition, Data curation. **L. Madhuanand:** Formal analysis, Data curation. **Y. Yogender:** Software, Formal analysis, Data curation. **B. Alsadik:** Supervision, Methodology, Formal analysis. **M. Weinmann:** Writing – review & editing, Validation, Resources, Methodology. **B. Jutzi:** Validation, Supervision, Resources, Data curation. **R. Qin:** Writing – review & editing, Software, Methodology, Formal analysis.

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.

Acknowledgements

The authors would like to thank ISPRS (International Society for Photogrammetry and Remote Sensing) for funding this benchmarking activity within the Scientific Initiative 2021 call. In addition, the authors thank AltoDrone s.r.l. for the data collection and support in the preprocessing phase.

References

Aanæs, H., Jensen, R.R., Vogiatzis, G., Tola, E., Dahl, A.B., 2016. Large-scale data for multiple-view stereopsis. Int. J. Comput. Vis. 120 (2), 153–168. https://doi.org/10.1007/s11263-016-0902-9.

Abayowa, B.O., Yilmaz, A., Hardie, R.C., 2015. Automatic registration of optical aerial imagery to a LiDAR point cloud for generation of city models. ISPRS J. Photogrammetry Remote Sens. 106, 68–81. https://doi.org/10.1016/j.isprsjprs.2015.05.006.

Badrinarayanan, V., Kendall, A., Cipolla, R. Segnet, 2017. A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39 (12), 2481–2495. https://doi.org/10.1109/TPAMI.2016.2644615.

Bleyer, M., Rhemann, C., Rother, C., 2011. Patchmatch stereo-stereo matching with slanted support windows. BMVC 11, 1–11. https://doi.org/10.5244/C.25.14.

Canny, J., 1986. A computational approach to edge detection. IEEE Trans. Pattern Anal. Mach. Intell. 8 (6), 679–698. https://doi.org/10.1109/TPAMI.1986.4767851.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2017. Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans. Pattern Anal. Mach. Intell. 40 (4), 834–848. https://doi.org/10.1109/TPAMI.2017.2699184.

Choy, C., Park, J., Koltun, V., 2019. Fully convolutional geometric features. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 8957–8965. https://doi.org/10.1109/ICCV.2019.00905. Seoul, Korea (South).

Dai, Y., Zhu, Z., Rao, Z., Li, B., 2019. MVS2: deep unsupervised multi-view stereo with multi-view symmetry. In: 2019 International Conference on 3D Vision (3DV). IEEE, pp. 1–8. https://doi.org/10.1109/3DV.2019.00010.

Dusmanu, M., Rocco, I., Pajdla, T., Pollefeys, M., Sivic, J., Torii, A., Sattler, T., 2019. D2-net: a trainable CNN for joint detection and description of local features. Proc. CVPR 8092–8101. https://doi.org/10.1109/CVPR.2019.00828.

Ebel, P., Mishchuk, A., Yi, K.M., Fua, P., Trulls, E., 2019. Beyond cartesian representations for local descriptors. ICCV 2019 253–262. https://doi.org/10.1109/ICCV.2019.00034.

Eigen, D., Fergus, R., 2015. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: Proceedings of the IEEE International Conference on Computer Vision. ICCV), pp. 2650–2658. https://doi.org/10.1109/ICCV.2015.304.

Eigen, D., Puhrsch, C., Fergus, R., 2014. Depth map prediction from a single image using a multi-scale deep network. Adv. Neural Inf. Process. Syst. 27.

Faugeras, O., Keriven, R., 1998. Complete dense stereovision using level set methods. In: European Conference on Computer Vision. Springer, pp. 379–393. https://doi.org/10.1007/BFb0055679.

Fu, H., Gong, M., Wang, C., Batmanghelich, K., Tao, D., 2018. Deep ordinal regression network for monocular depth estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2002–2011. https://doi.org/10.1109/CVPR.2018.00214.

Furukawa, Y., Curless, B., Seitz, S.M., Szeliski, R., 2010. Towards internet-scale multi-view stereo. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 1434–1441. https://doi.org/10.1109/CVPR.2010.5539802.

Galliani, S., Lasinger, K., Schindler, K., 2015. Massively parallel multiview stereopsis by surface normal diffusion. In: Proceedings of the *IEEE International Conference On Computer Vision*. ICCV), pp. 873–881. https://doi.org/10.1109/ICCV.2015.106.

Gallup, D., Frahm, J.-M., Mordohai, P., Yang, Q., Pollefeys, M., 2007. Real-time plane-sweeping stereo with multiple sweeping directions. In: 2007 *IEEE Conference On Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 1–8. https://doi.org/10.1109/CVPR.2007.383245.

Garg, R., Bg, V.K., Carneiro, G., Reid, I., 2016. Unsupervised CNN for single view depth estimation: geometry to the rescue. In: European Conference on Computer Vision (ECCV). Springer, pp. 740–756. https://doi.org/10.48550/arXiv.1603.04992.

Geiger, A., Lenz, P., Urtasun, R., 2012. Are we ready for autonomous driving? the KITTY vision benchmark suite. In: 2012 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 3354–3361. https://doi.org/10.1109/CVPR.2012.6248074.

Girshick, R., Donahue, J., Darrell, T., Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 580–587. https://doi.org/10.1109/CVPR.2014.81.

Glira, P., Pfeifer, N., Briese, C., Ressl, C., 2015a. Rigorous strip adjustment of airborne laser scanning data based on the ICP algorithm. In: ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences. Copernicus GmbH, pp. 73–80. https://doi.org/10.5194/isprsannals-II-3-W5-73-2015.

Glira, P., Pfeifer, N., Briese, C., Ressl, C.A., 2015b. Correspondence framework for ALS strip adjustments based on variants of the ICP algorithm. Photogramm. Fernerkund. GeoInf. 275–289. https://doi.org/10.1127/pfg/2015/0270.

Glira, P., Pfeifer, N., Mandlburger, G., 2019. Hybrid orientation of airborne LiDAR point clouds and aerial images. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences IV-2 (W5), 567–574. https://doi.org/10.5194/isprs-annals-IV-2-W5-567-2019.

Godard, C., Mac Aodha, O., Brostow, G.J., 2017. Unsupervised monocular depth estimation with left-right consistency. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 270–279. https://doi.org/10.1109/CVPR.2017.699.

Godard, C., Aodha, O. Mac, Firman, M., Brostow, G., 2019. Digging into self-supervised monocular depth estimation. In: Proceedings of the IEEE International Conference on Computer Vision. ICCV), pp. 3827–3837. https://doi.org/10.1109/ICCV.2019.00393.

Gonzalez-Aguilera, D., Ruiz de Ona, E., Lopez-Fernandez, L., Farella, E.M., Stathopoulou, E.K., Toschi, I., Remondino, F., Rodriguez-Gonzalvez, P., Hernandez-Lopez, D., Fusiello, A., Nex, F., 2020. Photomatch: an open-source multi-view and multi-modal feature matching tool for photogrammetric applications. ISPRS Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci. 43 (B5–2020), 213–219. https://doi.org/10.5194/isprs-archives-XLIII-B5-2020-213-2020.

Guo, X., Yang, K., Yang, W., Wang, X., Li, H., 2019. Group-wise correlation stereo network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3273–3282. https://doi.org/10.1109/CVPR.2019.00339.

Haala, N., Kölle, M., Cramer, M., Laupheimer, D., Mandlburger, G., Glira, P., 2020. Hybrid georeferencing, enhancement and classification of ultra-high resolution UAV LiDAR and image point clouds for monitoring applications. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences V-2, 727–734. https://doi.org/10.5194/isprs-annals-V-2-2020-727-2020, 2020.

Haala, N., Kölle, M., Cramer, M., Laupheimer, D., Zimmermann, F., 2022. Hybrid georeferencing of images and LiDAR data for UAV-based point cloud collection at millimetre accuracy. ISPRS Open Journal of Photogrammetry and Remote Sensing 4, 100014. https://doi.org/10.1016/j.ophoto.2022.100014.

Habib, A., Kim, E.M., Kim, C., 2007. New methodologies for true-orthophoto generation. Photogramm. Eng. Rem. Sens. 75 (1), 25–36. https://doi.org/10.14358/PERS.73.1.25.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778. https://doi.org/10.1109/CVPR.2016.90.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision. ICCV), pp. 2961–2969. https://doi.org/10.1109/ICCV.2017.322.

Hermann, M., Ruf, B., Weinmann, M., Hinz, S., 2020. Self-supervised learning for monocular depth estimation from aerial imagery. ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci. 357–364. https://doi.org/10.5194/isprs-annals-V-2-2020-357-2020. V-2–2020.

Hermann, M., Weinmann, M., Nex, F., Stathopoulou, E.K., Remondino, F., Jutzi, B., Ruf, B., 2024. Depth estimation and 3D reconstruction from UAV-borne imagery: evaluation on the UseGeo dataset. ISPRS Open Journal of Photogrammetry and Remote Sensing, 100065. https://doi.org/10.1016/j.ophoto.2024.100065. ISSN 2667-3932.

Hirschmüller, H., 2008. Stereo processing by semi-global matching and mutual information. IEEE Trans. Pattern Anal. Mach. Intell. 30 (2), 328–341. https://doi.org/10.1109/TPAMI.2007.1166.

Hosni, A., Bleyer, M., Rhemann, C., Gelautz, M., Rother, C., 2011. Real-time local stereo matching using guided image filtering. In: 2011 IEEE International Conference on Multimedia and Expo. IEEE, pp. 1–6. https://doi.org/10.1109/ICME.2011.6012131.

Hu, J., Ozay, M., Zhang, Y., Okatani, T., 2019. Revisiting single image depth estimation: toward higher resolution maps with accurate object boundaries. Proc. WACV 1043–1051. https://doi.org/10.1109/WACV.2019.00116.

Huang, P.-H., Matzen, K., Kopf, J., Ahuja, N., Huang, J.-B., 2018. DeepMVS: learning multi-view stereopsis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2821–2830. https://doi.org/10.1109/CVPR.2018.00298.

Im, S., Jeon, H.-G., Lin, S., Kweon, I.-S., 2019. Dpsnet: end-to-end deep plane sweep stereo. In: 7th International Conference on Learning Representations, pp. 1550–1554. https://doi.org/10.1109/LSP.2021.3099350.

Ji, M., Gall, J., Zheng, H., Liu, Y., Fang, L., 2017. Surfacenet: an end-to-end 3d neural network for multiview stereopsis. In: Proceedings of the IEEE International Conference on Computer Vision. ICCV), pp. 2307–2315. https://doi.org/10.1109/ICCV.2017.253.

Kaminski, R.S., Snavely, N., Seitz, S.M., Szeliski, R., 2009. Alignment of 3D point clouds to overhead images. In: *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 63–70. https://doi.org/10.1109/CVPRW.2009.5204180. Miami, FL, USA.

Kar, A., Häne, C., Malik, J., 2017. Learning a multi-view stereo machine. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 364–375. https://doi.org/10.5555/3294771.3294806.

Kendall, A., Martirosyan, H., Dasgupta, S., Henry, P., Kennedy, R., Bachrach, A., Bry, A., 2017. End-to-end learning of geometry and context for deep stereo regression. In: Proceedings of the IEEE International Conference on Computer Vision. ICCV), pp. 66–75. https://doi.org/10.1109/ICCV.2017.17.

Khot, T., Agrawal, S., Tulsiani, S., Mertz, C., Lucey, S., Hebert, M., 2019. Learning unsupervised multi-view stereopsis via robust photometric consistency. arXiv preprint arXiv 1905, 02706.

Knapitsch, A., Park, J., Zhou, Q.Y., Koltun, V., 2017. Tanks and temples: benchmarking large-scale scene reconstruction. ACM Trans. Graph. 36 (4), 1–13. https://doi.org/10.1145/3072959.3073599.

Kölle, M., Laupheimer, D., Schmohl, S., Haala, N., Rottensteiner, F., Wegner, J.D., Ledoux, H., 2021. The Hessigheim 3D (H3D) benchmark on semantic segmentation of high-resolution 3d point clouds and textured meshes from UAV lidar and multi-view-stereo. ISPRS Open Journal of Photogrammetry and Remote Sensing 1, 100001. https://doi.org/10.1016/j.ophoto.2021.100001.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. 25, 1097–1105. https://doi.org/10.1145/3065386.

Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., Navab, N., 2016. Deeper depth prediction with fully convolutional residual networks. In: 2016 Fourth International Conference on 3D Vision (3DV), pp. 239–248. https://doi.org/10.1109/3DV.2016.32.

Lee, J.H., Han, M.-K., Ko, D.W., Suh, I.H., 2019. From big to small: multi-scale local planar guidance for monocular depth estimation. arXiv preprint arXiv 1907, 10326. https://doi.org/10.48550/arXiv.1907.10326.

Li, J., Lee, G.H., 2021. DeepI2P: image-to-point cloud registration via deep classification. CVPR 2021 15960–156969. https://doi.org/10.1109/CVPR46437.2021.01570.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3431–3440. https://doi.org/10.1109/CVPR.2015.7298965.

Luo, Z., Zhou, L., Bai, X., Chen, H., Zhang, J., Yao, Y., Li, S., Fang, T., Quan, L., 2020. ASLFeat: learning local features of accurate shape and localization. Proc. CVPR 6588–6597. https://doi.org/10.1109/CVPR42600.2020.00662.

Lyu, Y., Vosselman, G., Xia, G.S., Yilmaz, A., Yang, M.Y., 2020. UAVid: a semantic segmentation dataset for UAV *imagery. ISPRS Journal of Photogrammetry and Remote Sensing*, 2020, 165, 108–119. https://doi.org/10.1016/j.isprsjprs.2020.05.009.

Madhuanand, L., Nex, F., Yang, M.Y., 2021. Self-supervised monocular depth estimation from oblique UAV videos. ISPRS J. Photogrammetry Remote Sens. 176, 1–14. https://doi.org/10.1016/j.isprsjprs.2020.05.009.

Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T., 2016. A large dataset to train convolutional networks for the disparity, optical flow, and scene flow estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4040–4048. https://doi.org/10.1109/CVPR.2016.438.

Merrell, P., Akbarzadeh, A., Wang, L., Mordohai, P., Frahm, J.-M., Yang, R., Nistér, D., Pollefeys, M., 2007. Real-time visibility-based fusion of depth maps. In: 2007 IEEE 11th International Conference on Computer Vision. IEEE, pp. 1–8. https://doi.org/10.1109/ICCV.2007.4408984.

Nex, F., Gerke, M., Remondino, F., Przybilla, H.J., Bäumker, M., Zurhorst, A., 2015. ISPRS benchmark for multi-platform photogrammetry. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences 2 (3), 135. https://doi.org/10.5194/isprsannals-II-3-W4-135-2015.

Nex, F., Armenakis, C., Cramer, M., Cucci, D.A., Gerke, M., Honkavaara, E., Kukko, A., Persello, C., Skaloud, J., 2022. UAV in the advent of the twenties: where we stand and what is next. ISPRS J. Photogrammetry Remote Sens. 184, 215–242. https://doi.org/10.1016/j.isprsjprs.2021.12.006.

Nex, F., Zhang, N., Remondino, F., Farella, E.M., Qin, R., Zhang, C., 2023. Benchmarking the extraction of 3D geometry from UAV images with deep learning methods. Int. Arch. Photogram. Rem. Sens. Spatial Inf. Sci. 48, 123–130. https://doi.org/10.5194/isprs-archives-XLVIII-1-W3-2023-123-2023.

Paschalidou, D., Ulusoy, O., Schmitt, C., Van Gool, L., Geiger, A. Raynet, 2018. Learning volumetric 3d reconstruction with ray potentials. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR), pp. 3897–3906. https://doi.org/10.1109/CVPR.2018.00410.

Qin, R.. Multi-view stereo processor. Available: https://u.osu.edu/qin.324/msp/. (Accessed 1 August 2023). accessed.

Remondino, F., Spera, M.G., Nocerino, E., Menna, F., Nex, F., 2014. State of the art in high-density image matching. Photogramm. Rec. 29 (146), 144–166. https://doi.org/10.1111/phor.12063.

Remondino, F., Menna, F., Morelli, L., 2021. Evaluating hand-crafted and learning-based features for photogrammetric applications. In: *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Volume XLIII-B2-2021, XXIV ISPRS Congress, 2021 edition. https://doi.org/10.5194/isprs-archives-XLIII-B2-2021-549-2021.

Remondino, F., Morelli, L., Stathopoulou, E., Elhashash, Qin, M., 2022. Aerial triangulation with learning-based tiepoints. Int. Arch. Photogram. Rem. Sens. Spatial Inf. Sci. XLIII-B2–2022 https://doi.org/10.5194/isprs-archives-XLIII-B2-2022-77-2022. XXIV ISPRS Congress (2022 edition).

Remondino, F., Karami, A., Yan, Z., Mazzacca, G., Rigon, S., Qin, R., 2023. A critical analysis of NeRF-based 3D reconstruction. Rem. Sens. 15, 3585.

Revaud, J., Weinzaepfel, P., de Souza, C.R., Humenberger, M., 2019. R2D2: repeatable and reliable detector and descriptor. Proc. NIPS 2019 12414–12424. https://doi.org/10.48550/arXiv.1906.06195, 2019.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference Proceedings, Part III. Springer International Publishing, pp. 234–241. https://doi.org/10.1007/978-3-319-24574-4_28, 18.

Rotstein, N., Bracha, A., Kimmel, R., 2022. Multimodal coloured point cloud to image alignment. CVPR 2022 6656–6666. https://doi.org/10.1109/CVPR52688.2022.00654.

Savinov, N., Seki, A., Ladicky, L., Sattler, T., Pollefeys, M., 2017. Quad-networks: unsupervised learning to rank for interest point detection. Proc. CVPR 3929–3937. https://doi.org/10.1109/CVPR.2017.418, 2017.

Saxena, A., Sun, M., Ng, A.Y., 2008. Make3d: learning 3d scene structure from a single still image. IEEE Trans. Pattern Anal. Mach. Intell. 31 (5), 824–840. https://doi.org/10.1109/TPAMI.2008.132.

Scharstein, D., 1994. Matching images by comparing their gradient fields. In: Proceedings of the 12th International Conference on Pattern Recognition, 1. IEEE, pp. 572–575.

Schönberger, J.L., Frahm, J.-M., 2016b. Structure-from-motion revisited. In: Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR), pp. 4104–4113. https://doi.org/10.1109/CVPR.2016.445.

Schönberger, J.L., Zheng, E., Frahm, J.-M., Pollefeys, M., 2016. Pixelwise view selection for unstructured multi-view stereo. In: Proceedings of the European Conference on Computer Vision (ECCV). Springer, pp. 501–518. https://doi.org/10.1007/978-3-319-46487-9_31.

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: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3260–3269. https://doi.org/10.1109/CVPR.2017.272.

Seitz, S.M., Curless, B., Diebel, J., Scharstein, D., Szeliski, R., 2006. A comparison and evaluation of multi-view stereo reconstruction algorithms. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 1. IEEE, pp. 519–528. https://doi.org/10.1109/CVPR.2006.19.

Silberman, N., Hoiem, D., Kohli, P., Fergus, R., 2012. Indoor segmentation and support inference from RGBD images. In: European Conference on Computer Vision. Springer, pp. 746–760. https://doi.org/10.1007/978-3-642-33715-4_54.

Stathopoulou, E.K., Remondino, F., 2023. A survey of conventional and learning-based methods for multi-view stereo. Photogramm. Rec. https://doi.org/10.1111/phor.12456.

Strecha, C., Fransens, R., Van Gool, L., 2004. Wide-baseline stereo from multiple views: a probabilistic account. In: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 1. IEEE, p. I. https://doi.org/10.1109/CVPR.2004.1315080.

Strecha, C., Fransens, R., Van Gool, L., 2006. Combined depth and outlier estimation in multi-view stereo. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2. IEEE, pp. 2394–2401. https://doi.org/10.1109/CVPR.2006.78.

Teed, Z., Deng, J., 2019. DeepV2D: video to depth with differentiable structure from motion. In: International Conference on Learning Representations. https://doi.org/10.48550/arXiv.1812.04605.

Toschi, I., Farella, E.M., Welponer, M., Remondino, F., 2021. Quality-based registration refinement of airborne LiDAR and photogrammetric point clouds. ISPRS J. Photogrammetry Remote Sens. 172, 160–170. https://doi.org/10.1016/j.isprsjprs.2020.12.005.

Tosi, F., Aleotti, F., Poggi, M., Mattoccia, S., 2019. Learning monocular depth estimation infusing traditional stereo knowledge. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9799–9809. https://doi.org/10.1109/CVPR.2019.01003.

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. https://doi.org/10.1016/j.displa.2021.102102.

Watson, J., Firman, M., Brostow, G.J., Turmukhambetov, D., 2019. Self-supervised monocular depth hints. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. CVPR), pp. 2162–2171. https://doi.org/10.1109/ICCV.2019.00225.

Welponer, M., Stathopoulou, E.-K., Remondino, F., 2022. Monocular depth prediction in photogrammetric applications. Int. Arch. Photogram. Rem. Sens. Spatial Inf. Sci. 43, 469–476. https://doi.org/10.5194/isprs-archives-XLIII-B2-2022-469-2022.

Wenzel, K., Rothermel, M., Haala, N., Fritsch, D., 2013. Sure–the IfP software for dense image matching. Photogrammetric Week 13, 59–70.

Wu, T., Vallet, B., Pierrot-Deseilligny, M., Rupnik, E., 2021. A new stereo dense matching benchmark dataset for deep learning. Int. Arch. Photogram. Rem. Sens. Spatial Inf. Sci. 405–412. https://doi.org/10.5194/isprs-archives-XLIII-B2-2021-405-2021. XLIII-B2-2021.

Xu, Q., Tao, W., 2019. Multi-scale geometric consistency guided multi-view stereo. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR), pp. 5483–5492. https://doi.org/10.1109/CVPR.2019.00563.

Xu, Q., Tao, W., 2020. Learning inverse depth regression for multi-view stereo with correlation cost volume. Proc. AAAI Conf. Artif. Intell. 34, 12508–12515. https://doi.org/10.1609/aaai.v34i07.6939.

Xu, D., Wang, W., Tang, H., Liu, H., Sebe, N., Ricci, E., 2018. Structured attention guided convolutional neural fields for monocular depth estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3917–3925. https://doi.org/10.1109/CVPR.2018.00412.

Xu, Q., Oswald, M.R., Tao, W., Pollefeys, M., Cui, Z., 2021. Non-local recurrent regularization networks for multi-view stereo. arXiv preprint arXiv:2110.06436. https://doi.org/10.48550/arXiv.2110.06436.

Xu, N., Qin, R., Song, S., 2023. Point cloud registration for LiDAR and photogrammetric data: a critical synthesis and performance analysis on classic and deep learning algorithms. ISPRS Open Journal of Photogrammetry and Remote Sensing 8, 100032. https://doi.org/10.1016/j.ophoto.2023.100032. ISSN 2667-3932.

Yadav, Y., Alsadik, B., Nex, F., Remondino, F., Glira, P., 2023. Hybrid adjustment of UAS-based lidar and image data. Int. Arch. Photogram. Rem. Sens. Spatial Inf. Sci. 633–640. https://doi.org/10.5194/isprs-archives-XLVIII-1-W2-2023-633-2023. XLVIII-1/W2-2023.

Yan, S., Zhang, M., Peng, Y., Liu, Y., Tan, H., 2022. AgentI2P: optimizing image-to-point cloud registration via behaviour cloning and reinforcement learning. Rem. Sens. 14, 6301. https://doi.org/10.3390/rs14246301.

Yang, B., Chen, C., 2015. Automatic registration of UAV-borne sequent images and LiDAR data. ISPRS J. Photogrammetry Remote Sens. 101, 262–274. https://doi.org/10.1016/j.isprsjprs.2014.12.025.

Yang, J., Mao, W., Alvarez, J.M., Liu, M., 2020. Cost volume pyramid based depth inference for multi-view stereo. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4877–4886. https://doi.org/10.1109/TPAMI.2021.3082562.

Yao, Y., Luo, Z., Li, S., Fang, T., Quan, L., 2018. MVSnet: depth inference for unstructured multi-view stereo. In: Proceedings of the European Conference on Computer Vision. ECCV), pp. 767–783. https://doi.org/10.1007/978-3-030-01237-3_47.

Yao, Y., Luo, Z., Li, S., Shen, T., Fang, T., Quan, L., 2019. Recurrent MVSnet for high-resolution multi-view stereo depth inference. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR), pp. 5525–5534. https://doi.org/10.1109/CVPR.2019.00567.

Yao, Y., Luo, Z., Li, S., Zhang, J., Ren, Y., Zhou, L., Fang, T., Quan, L.BlendedMVS., 2020. A large-scale dataset for generalized multi-view stereo networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1790–1799. https://doi.org/10.1109/CVPR42600.2020.00186.

Yin, W., Liu, Y., Shen, C., Yan, Y., 2019. Enforcing geometric constraints of virtual normal for depth prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. CVPR), pp. 5684–5693. https://doi.org/10.1109/ICCV.2019.00578.

Yin, W., Wang, X., Shen, C., Liu, Y., Tian, Z., Xu, S., Sun, C., Renyin, D., 2020. Diversedepth: affine-invariant depth prediction using diverse data. arXiv preprint arXiv:2002.00569. https://doi.org/10.48550/arXiv.2002.00569.

Yin, W., Zhang, J., Wang, O., Niklaus, S., Mai, L., Chen, S., Shen, C., 2021. Learning to recover 3d scene shape from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 204–213. https://doi.org/10.1109/CVPR46437.2021.00027.

Zbontar, J., LeCun, Y., 2015. Computing the stereo matching cost with a convolutional neural network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1592–1599. https://doi.org/10.1109/CVPR.2015.7298767.

Zhang, Z., Dai, Y., Sun, J., 2020. Deep learning based point cloud registration: an overview. Virtual Reality & Intelligent Hardware 2 (3), 222–246. https://doi.org/10.1016/j.vrih.2020.05.002.

Zhang, N., Nex, F., Vosselman, G., Kerle, N., 2023. Lite-Mono: a lightweight CNN and Transformer architecture for self-supervised monocular depth estimation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 18537–18546.

Zhou, T., Brown, M., Snavely, N., Lowe, D.G., 2017. Unsupervised learning of depth and ego-motion from video. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6612–6619. https://doi.org/10.1109/CVPR.2017.700.

Zhou, K., Meng, X., Cheng, B., 2020. Review of stereo matching algorithms based on deep learning. Comput. Intell. Neurosci. https://doi.org/10.1155/2020/8562323.

Zhou, T., Hasheminasab, S.M., Habib, A., 2021. Tightly-coupled camera/LiDAR integration for point cloud generation from GNSS/INS-assisted UAV mapping systems. ISPRS J. Photogrammetry Remote Sens. 180, 336–356. https://doi.org/10.1016/j.isprsjprs.2021.08.020.