ISPRS Journal of Photogrammetry and Remote Sensing 198 (2023) 84–98

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

ENRICH: Multi-purposE dataset for beNchmaRking In Computer vision and pHotogrammetry

Davide Marelli [a,*], Luca Morelli [b,c], Elisa Mariarosaria Farella [b], Simone Bianco [a], Gianluigi Ciocca [a], Fabio Remondino [b]

[a] Imaging and Vision Laboratory, Department of Informatics, Systems and Communication, University of Milano-Bicocca, Viale Sarca 336, Milano, 20126, Italy
[b] 3D Optical Metrology (3DOM) unit, Bruno Kessler Foundation (FBK), Via Sommarive 18, Trento, 38123, Italy
[c] Department of Civil, Environmental and Mechanical Engineering (DICAM), University of Trento, Via Mesiano 77, Trento, 38123, Italy

ARTICLE INFO

Keywords:
Dataset
Image matching
Photogrammetry
Local features
3D reconstruction
Depth estimation
Synthetic images

ABSTRACT

The availability of high-resolution data and accurate ground truth is essential to evaluate and compare methods and algorithms properly. Moreover, it is often difficult to acquire real data for a given application domain that is sufficiently representative and heterogeneous in terms of scene representation, acquisition conditions, point of view, etc. To overcome the limitations of available datasets, this paper presents a new synthetic, multi-purpose dataset called ENRICH for testing photogrammetric and computer vision algorithms. Compared to existing datasets, ENRICH offers higher resolution images rendered with different lighting conditions, camera orientations, scales, and fields of view. Specifically, ENRICH is composed of three sub-datasets: ENRICH-Aerial, ENRICH-Square, and ENRICH-Statue, each exhibiting different characteristics. We show the usefulness of the proposed dataset on several examples of photogrammetry and computer vision-related tasks such as: evaluation of hand-crafted and deep learning-based local features, effects of ground control points (GCPs) configuration on the 3D accuracy, and monocular depth estimation. We make ENRICH publicly available at: https://github.com/davidemarelli/ENRICH.

1. Introduction

Over the past two decades, the need to evaluate new computer vision and photogrammetry algorithms has motivated the research community toward the creation of 2D and 3D datasets for different scenarios (e.g., indoor, outdoor, laboratory, urban, buildings) and tasks (such as image matching, image retrieval, structure-from-motion, and SLAM).

As a tool for scientists, a benchmark dataset is a set of data to evaluate or compare the performance of sensors, platforms, or processing algorithms (Bakula et al., 2019) against a high-quality and accurate ground truth, although the acquisition of a sufficient amount of data is still a challenge. Barriers to their realization are related to the costs and time for obtaining data diversity (like scale or scene), and the collection of precise annotations and accurate and reliable ground truth.

Driven by the achievements and pending research issues in the 3D reconstruction sector (Remondino et al., 2021; Bellavia et al., 2022a), many scientific initiatives have been recently proposed to evaluate the current status of available processing methods while boosting further investigations, both in the photogrammetric and computer vision research fields. These activities have encouraged developers and users to deliver comparative performance analyses focusing, in particular, on image-based 3D reconstruction (Schonberger et al., 2017; Bianco et al., 2018; Jin et al., 2021).

In recent years, advanced deep learning-based algorithms for extracting complex information and features from visual data have been effectively applied in many domains and scenarios, such as image analysis, remote sensing, computer vision, and geoscience research, outperforming standard approaches or opening to new applications not possible with classical methods. However, applying these techniques implies the availability of a large amount of data for training the underlying models (LeCun et al., 2015). Moreover, real data collected and targeted for a given task are rarely complete and heterogeneous enough (in terms of, e.g., acquisition condition, point of view, signal distortion) to enable the design of robust and flexible algorithms and approaches. Finally, the training set must be annotated, and this process is frequently time-consuming and resource-intensive. Synthetic heterogeneous and annotated datasets have proved to be an effective and

* Corresponding author.
E-mail addresses: davide.marelli@unimib.it (D. Marelli), lmorelli@fbk.eu (L. Morelli), elifarella@fbk.eu (E.M. Farella), simone.bianco@unimib.it (S. Bianco), gianluigi.ciocca@unimib.it (G. Ciocca), remondino@fbk.eu (F. Remondino).

https://doi.org/10.1016/j.isprsjprs.2023.03.002
Received 26 July 2022; Received in revised form 15 February 2023; Accepted 3 March 2023
Available online 10 March 2023
0924-2716/© 2023 The Authors. Published by Elsevier B.V. on behalf of International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS). This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Table 1
Comparison of the most popular benchmark datasets with the proposed ENRICH dataset.

Name	Setting	Real/Synthetic	Resolution	Image format	Videos	SfM	SLAM	Stereo	MVS	Depth	GCP	Lighting
Middleburry stereo	Indoor	Real	2–6 Mpx	Fixed	–	–	–	✓	–	✓	–	Varying
Middleburry MVS	Lab	Real	0.3 Mpx	Fixed	–	✓	–	–	✓	–	–	Fixed
DTU	Lab	Real	2 Mpx	Fixed	–	✓	–	–	✓	–	–	Varying
KITTI	Streets	Real	0.5 Mpx	Fixed	✓	–	✓	✓	✓	✓	–	Fixed
Strecha	Monuments	Real	6 Mpx	Fixed	–	✓	–	–	✓	–	–	Fixed
Tanks and Temples	Indoor/Outdoor	Real	8 Mpx	Fixed	✓	✓	✓	–	✓	–	–	Fixed
3DOMcity	Lab	Real	24 Mpx	Varying	–	✓	–	–	✓	–	✓	Fixed
ETH3D	Indoor/Outdoor	Real	0.4–24 Mpx	Varying	✓	✓	✓	✓	✓	✓	–	Fixed
DIODE	Indoor/Outdoor	Real	0.8 Mpx	Fixed	–	–	–	–	–	✓	–	Varying
IVL-SYNTHSFM-v2	Objects	Synthetic	2 Mpx	Varying	–	✓	–	–	✓	–	–	Varying
BlendedMVS	Multi-scale	Blended	3 Mpx	Fixed	–	✓	–	–	✓	✓	–	Varying
ENRICH	Multi-scale outdoor	Synthetic	24 Mpx	Varying	–	✓	–	–	✓	✓	✓	Varying

efficient way to overcome the current limitations of real data (Tremblay et al., 2018; Yao et al., 2020; Nikolenko et al., 2021).

This paper introduces ENRICH, a new multi-purpose synthetic benchmark dataset created to (i) complement existing close-range and aerial datasets, (ii) boost investigations and analyses in the 3D reconstruction process, and (iii) evaluate photogrammetric and computer vision algorithms. To this end, the dataset comprises three collections of outdoor images capturing an urban area taken from above, a city square, and a statue. Compared to existing benchmarks datasets, ENRICH offers higher-resolution images that are rendered with different lighting conditions (clear blue sky, cloudy sky, sunrise, etc.) and camera orientation (landscape and portrait). ENRICH can be exploited to benchmark algorithms under variable and realistic conditions with data acquired at diverse scales, different cameras, and lighting setups.

Unlike many benchmarks and challenges, such as the Image Matching Challenge (Jin et al., 2021), we have included in the dataset images with rotations from 0 to 180 degrees to encourage the scientific community to propose solutions that also manage rotations, a property of local features that is fundamental in many photogrammetric applications.

The contribution of the paper is twofold:

- to introduce the ENRICH dataset and its characteristics, consisting of three 2D and 3D synthetic and multi-scale outdoor datasets: an urban area (ENRICH-Aerial), a square (ENRICH-Square), and a statue (ENRICH-Statue). The benchmark includes, for the first time, high-resolution rendered images, depth maps, camera parameters, absolute orientation information, Ground Control Points (GCPs) and 3D models as ground truth data, allowing multiple investigations in the fields of photogrammetry and computer vision.
- to show the usefulness of the ENRICH dataset by performing several processing tests exploring some steps of the photogrammetric 3D reconstruction pipeline. We propose three example analyses to be addressed with the ENRICH datasets: (i) the contribution of new deep learning-based local features for image matching and the evaluation of different structure-from-motion (SfM) pipelines; (ii) the influence of GCPs number and distribution in aerial mapping applications; (iii) the use of neural networks for monocular depth estimation.

The organization of the paper is as follows: Section 2 gives an overview of the state-of-the-art of existing benchmark datasets and the tasks covered, Section 3 describes the design and creation of the ENRICH dataset, and Section 4 presents three application scenarios evaluated on the ENRICH data. Finally, in Section 5 we report our conclusions.

2. Related work

This section presents the most popular public datasets and benchmark datasets available for the photogrammetric and computer vision research communities (Fig. 1). Table 1 summarizes their characteristics compared to the ENRICH dataset. For each dataset we report the acquisition setting, its source, image properties, covered tasks, and lighting conditions.

- **Middleburry stereo**[1] (Scharstein et al., 2014). The dataset contains 32 + 24 stereo scenes (published in 2014 and 2021, respectively) to evaluate stereo algorithms on 6 and 2-megapixel images. Each scene is composed of a single stereo pair with substantial exposure variations, while scenes and cameras are static. The ground truth consists of accurate depth maps.
- **Middleburry MVS**[2] (Seitz et al., 2006). The dataset, designed for evaluating multi-view stereo reconstruction algorithms, consists of undistorted images (640 × 480 pixels) of a plaster Greek temple and a dinosaur. The ground truth is a laser-scanner model (not released), while image orientations are provided.
- **DTU Robot Image Data Sets**.[3] It contains two different collections of scenes, one designed to evaluate local features (Aanæs et al., 2012), the other for multi-view-stereo (MVS) investigations (Jensen et al., 2014). Images (2 megapixels) of miniatures were acquired, varying the illumination conditions. A structured light scanner was mounted on the same arm, and both the camera poses and the geometric model are provided. The trajectory of the images follows a circular path for all objects.
- **KITTI**[4] (Geiger et al., 2012). This benchmark dataset is used in the context of autonomous driving. The data derive from several devices mounted on a car: two grayscale and two color cameras, a laser scanner, and an inertial navigation system (GPS/IMU). The goal was to provide training and testing images for different computer vision tasks, such as stereo matching, SLAM, 3D object detection, and depth prediction.
- **Strecha**[5] (Strecha et al., 2008). This benchmark dataset is designed to compare the reliability of passive 3D reconstruction methods with active stereo systems. Data consists of LiDAR and camera acquisitions of outdoor scenes up to 6 megapixels. The authors have provided camera poses and calibrations, and the laser-scanner model.
- **Tanks and Temples**[6] (Knapitsch et al., 2017). It is a benchmark dataset for image-based 3D reconstruction algorithms. The data consists of real outdoor scenes divided into training and test sets derived from 4K videos (acquired with two rolling shutter and one global-shutter camera). Laser scanner point clouds are provided as ground truth, as well as the reconstruction and camera poses obtained by processing the images with an "out of the box" approach based on COLMAP (Schonberger and Frahm, 2016).

[1] https://vision.middlebury.edu/stereo/data/
[2] https://vision.middlebury.edu/mview/
[3] http://roboimagedata.compute.dtu.dk/
[4] http://www.cvlibs.net/datasets/kitti/
[5] http://cvlab.epfl.ch/data
[6] https://www.tanksandtemples.org/

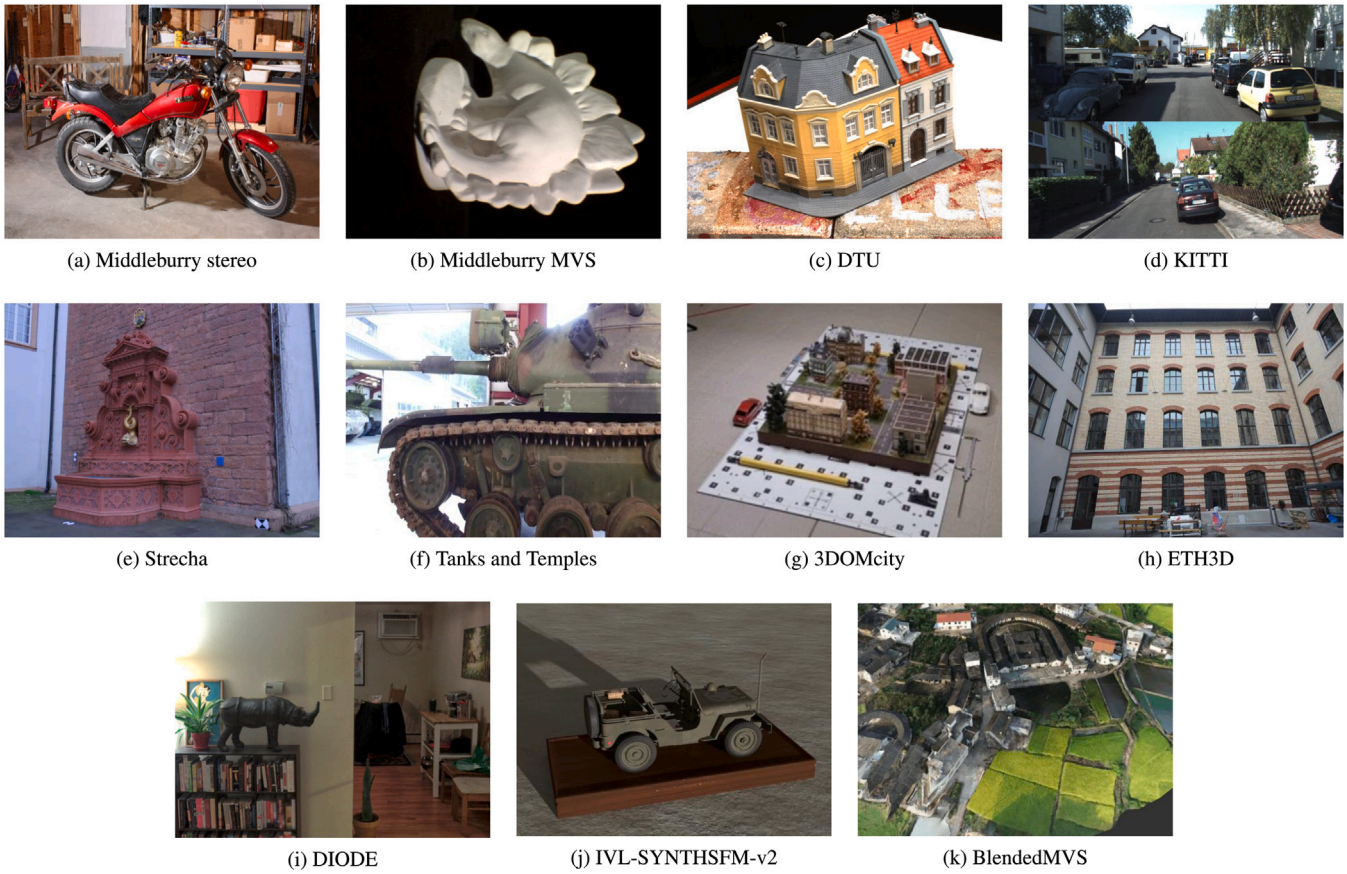


Fig. 1. Samples from the most popular benchmark datasets reported in Table 1.

- **3DOMcity**[7] (Özdemir et al., 2019). It is a multipurpose and high-resolution (6016 × 4016 pixels) benchmark dataset, including 420 nadir and oblique aerial images, to assess the performance of the image-based pipeline for 3D urban reconstruction and 3D data classification. Laser scanner point clouds and reference measurements are provided to evaluate image orientation, dense image matching, and point cloud classification results.
- **ETH3D**[8] (Schops et al., 2017). This dataset is composed of multi-sensors low and high-resolution images and videos for MVS investigations, with scenes acquired both indoor and outdoor, and laser-scanner data as ground truth. Moreover, a benchmark is available dedicated to the evaluation of SLAM algorithms.
- **DIODE**[9] (Vasiljevic et al., 2019). It is a dataset specifically created for the depth estimation task. It contains color images with accurate laser-scanned depth measurements of indoor and outdoor scenes. It also includes validity masks for the ground truth depth scans and surfaces normal vector ground truth.
- **IVL-SYTHNSFM-v2**[10] (Marelli et al., 2020). This dataset comprises 4000 synthetic images, taken from 100 points of view and portraying five 3D models. For each model, 8 scenes with different lighting combinations, depth of field, and motion blur are created. Data also includes information about each image's intrinsic and extrinsic parameters and 3D models as ground truth.
- **BlendedMVS**[11] (Yao et al., 2020). Similarly to ENRICH, it provides synthetic sets of images consisting of about 17,000 rendered images with a max resolution of 2048 × 1536 pixels representing 113 different scenes, both aerial and terrestrial. The peculiarity of this dataset is that the rendered images are obtained from a linear combination of low-pass and high-pass filters applied to the original and rendered images to preserve the realism of lights. This procedure implies that no new synthetic views can be generated in the dataset. BlendedMVG is a superset of BlendedMVS, expanded with additional 389 scenes for a total of about 110,000 rendered images.

The tasks covered by the presented benchmark datasets are mainly related to evaluating the performance of the entire SfM pipeline, focusing, on multi-view stereo reconstruction algorithms (Middleburry MVS, DTU, Strecha, KITTI, 3DOM city, Tanks and Temples, ETH3D, and IVL-SYNTHSFM-v2). In addition, KITTI and ETH3D propose methods for evaluating SLAM or Visual Odometry techniques. The strength of KITTI is the long image sequences for autonomous driving applications. Tanks and Temples use only video sequences for SfM reconstructions, while Strecha offers, in addition to standard datasets with calibrated cameras, also datasets with uncalibrated cameras, e.g., for reconstructions from internet photos. DIODE and BlendedMVS are designed to train algorithms for depth estimation. KITTI also allows evaluating algorithms for object detection, while 3DOMcity deals with classification problems. IVL-SYNTHSFM-v2 and BlendedMVS are the only datasets with accurate ground truth from synthetic models and image generation. Middleburry stereo, DTU, DIODE, IVL-SYNTHSFM-v2, and BlendedMVS present environments with very different lighting conditions within the same scene to test their effects on image orientation and final 3D reconstruction.

The ENRICH datasets we present are intended to complement existing benchmark datasets, most of which offer low-resolution images and/or scenes of limited size. BlendedMVS also proposes multi-scale

[7] https://3dom.fbk.eu/3domcity-benchmark
[8] http://www.eth3d.net
[9] https://diode-dataset.org/
[10] https://board.unimib.it/datasets/fnxy8z8894
[11] https://github.com/YoYo000/BlendedMVS

Table 2
Summary of the acquisition setup of each dataset in our ENRICH benchmark.

Dataset	Camera	Focal Length	Images	Orientation	Lighting setup	GSD
ENRICH-Aerial	Nadir	35 mm/5882 px	60	Landscape	Uniform light	2.5 cm
	Forward	70 mm/11,764 px	60	Landscape	Uniform light	1.8 cm
	Backward	70 mm/11,764 px	60	Landscape	Uniform light	1.8 cm
	Right	70 mm/11,764 px	60	Landscape	Uniform light	1.8 cm
	Left	70 mm/11,764 px	60	Landscape	Uniform light	1.8 cm
	Nadir 2	35 mm/5,882 px	39	Landscape	Uniform light	3.0 cm
ENRICH-Square	Camera 1	35 mm/5882 px	50	Landscape & Portrait	Partly cloudy	0.8 cm
	Camera 2	50 mm/8403 px	50	Landscape	Clear sky	0.5 cm
	Camera 3	35 mm/5882 px	50	Landscape & Portrait	Sunrise	0.8 cm
	Camera 4	35 mm/5882 px	50	Landscape	Clear sky	1.0 cm
ENRICH-Statue	Camera 1	50 mm/8403 px	50	Landscape	Partly cloudy	0.69 mm
	Camera 2	35 mm/5882 px	50	Portrait	Clear sky	0.64 mm
	Camera 3	50 mm/8403 px	50	Landscape	Sunrise	0.70 mm
	Camera 4	35 mm/5882 px	50	Portrait	Cloudy	0.64 mm

datasets, but with quite low-resolution images and only upright. ENRICH datasets jointly present high-resolution and multi-scale images, the inclusion of camera rotations, and accurate ground truth of poses and 3D models. Accuracy is ensured by synthetic image generation, while the use of 3D reality-based surveys and 3D synthetic models guarantees texture realism.

3. The ENRICH dataset

The ENRICH benchmark dataset consists of three synthetic datasets generated from 3D scenes reproducing different scenarios, levels of detail, resolution, scale, lighting conditions, and fields of view (FoV): ENRICH-Aerial, ENRICH-Square, and ENRICH-Statue. The 3D models used to compose the scenes come from real objects, assuring realistic textures. In the case of the ENRICH-Square scene, the trees in were included to add more details. Attention have been paid to ensure realistic geometry and texture.

The ENRICH-Aerial dataset is generated from an aerial image block of the city of Launceston, Australia (Launceston City Council, 2017). The ENRICH-Square and the ENRICH-Statue are two ground-level datasets, capturing, respectively, a square surrounded by monumental buildings and a statue placed in its center. A virtual camera, based on the specifications of the Nikon D750 DSLR full-frame camera (sensor size 35.9 × 24 mm, pixel size 5.95 μm) with an image resolution of 6016 × 4016 px (24 MP) is used to acquire images of the scenes, and create the corresponding datasets. The virtual camera then acquires ideal images without lens distortions (pinhole camera model).

In all the scenes, GCPs were manually inserted on flat areas and well distributed at different heights. An overview of the acquisition setup for each one of the three datasets is given in Table 2.

To create these datasets, we leverage our previous experiences in generating synthetic datasets using computer graphics methods and tools. In Bianco et al. (2018), we developed a Blender plug-in to generate a dataset of different objects acquired with a virtual camera under different lighting and pose conditions (Marelli et al., 2022). The created dataset (Marelli et al., 2020) was used to evaluate several structure-from-motion pipelines with pixel-precise ground truths. Here we applied our previous knowledge to generate the new datasets by integrating new functionalities in our previously developed tools specifically tailored for the multi-scale tasks. In particular new functionalities have been introduced to support multiple camera paths and configurations, automatic export of GCP coordinates and image visibility, and depth data.

Compared to our previous synthetic dataset, IVL-SYNTHSFM, created mainly to evaluate SfM pipelines, ENRICH offers a more diverse set of scenes, acquired with a variable range of cameras and scales. This makes ENRICH a multi-purpose dataset, exploitable for different photogrammetry and computer vision tasks, including SfM applications.

Additional details on 3D scenes, acquisition, and data available for each dataset are provided in the following sections.

Fig. 2. Orthographic view of the GCP placement on the ENRICH-Aerial dataset. Cross and round shapes as in Fig. 3.

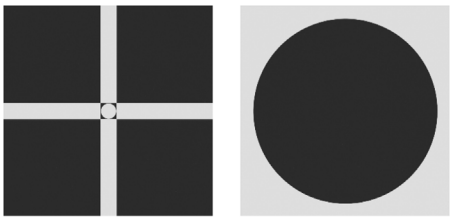
3.1. Data generation method

Data were generated using the popular 3D modeling and rendering software Blender (Blender Online Community, 2018) with the aid of the SfM Flow add-on (Marelli et al., 2022). This add-on has been extended to support multi-camera configurations as well as GCP placement and export of their ground truth 3D position and visibility in the images.

For the ENRICH-Aerial dataset, a Blender scene was created importing the 3D mesh and textures of a 3D LIDAR scan of the city of Launceston (Launceston City Council, 2017). The original scene is composed of different tiles at different LOD. Each tile is about 200 m × 200 m square. The capture pixel size is 2–3 cm and 5 cm GSD. The spatial accuracy is 0.05 m (absolute accuracy in XYZ).[12] A total of 26 GCPs of 50 cm × 50 cm square are positioned in the scene on flat or almost-flat surfaces at different elevations (Fig. 2), with a cross (see Fig. 3a) or a circular pattern (see Fig. 3b). The size of each GCP is defined to have the cross's thickness visible in at least 4 pixels. Each GCP is guaranteed to be visible in at least ten images. We chose to use this GCPs configuration in order to have them distributed uniformly in the scene and visible in many images. This allows researchers to select which GCPs to use for their experiments and still have targets available to be used as Check Points (see Section 4.2).

The acquisition is performed simulating a typical oblique aerial camera with five views (see Fig. 4a–e): one nadir and four oblique views (forward, backward, left, and right). The nadir camera has a focal length of 35 mm, whereas that of the oblique cameras is 70 mm. The oblique cameras have an angle of 45° w.r.t. the nadir direction. The

[12] https://github.com/stuarta0/launceston-3d



(a) Cross shape (b) Round shape

Fig. 3. Ground Control Point patterns.

five cameras are rigidly mounted on a virtual flying platform at the same altitude, with the oblique ones having a 20 cm padding from the nadir camera in their viewing direction. The image acquisitions followed six parallel strips, with 10 acquisition points in each track, providing a total of 300 images. A second acquisition, orthogonal to the first one, is performed using only the nadir camera (nadir-2). This path consist of three parallel strips with 13 images each, for a total of 39 images. In both paths the image overlap for the nadir images is 80% along the track and 60% across it, respectively. The flying heights are approximately 150 m and 175 m above the ground. These camera paths mimic real acquisition setups that allow detailed 3D model generation (Rupnik et al., 2015). The images taken by the first nadir camera have an average Ground Sample Distance (GSD) of 2.5 cm, while for the second nadir camera they have a GSD of 3.0 cm. The GSD of the images of the oblique cameras ranges from 1.2 to 2.4 cm (1.8 cm at an average distance of 213 m). The paths followed by the cameras are visible in Fig. 5. In the ENRICH-Aerial dataset, the scene is illuminated only by a global environment white light since the diffuse texture of the 3D Launceston model already incorporates shadows. The shadows are embedded in the model due to the use of multiple images to build it. Blending of those images, acquired at different times, produced some artifacts in the textures. While this has no impact on the evaluation of geometry-related tasks, it may limit the usability of ENRICH-Aerial on texture-related tasks, such as image blending, shadow removal, and de-lighting. The images were generated employing the Blender's Eevee raster render engine (Blender Online Community, 2021) that focuses on rendering speed while achieving Physically Based Rendering of materials.

In the ENRICH-Square dataset, several 3D models from different sources were used to build the virtual scene (for more details see the readme file and the source links accompanying the dataset). It comprises 3D meshes of monumental buildings surrounding the square, statues, and trees. The tallest building is 27 m high. The meshes of the buildings were generated using photogrammetry software such as Agisoft Metashape by the original authors; trees and walls were instead subsequently added to provide a more detailed and cohesive scene. Please refer to the readme file included in the dataset for a complete list of models and methods used to create them. In some cases, 3D model editing was required to solve geometry issues in the meshes (e.g. holes on the facades due to occlusions or dark areas). The whole square is surrounded by a hilly landscape model, providing a background for some far portions of the scene (i.e., behind the walls). Cross pattern GCPs of size $15\ \text{cm} \times 15\ \text{cm}$ square are positioned on the facades of the buildings at different heights (see Fig. 6); a total of 54 GCPs are available, and each one is visible in at least 16 images. This GCP placement guarantees that they are uniformly distributed in the scene and visible in a large amount of images.

Images are captured by four cameras (Fig. 7a–d), each of them providing 50 images for a total of 200. Images for Camera1 are acquired following a circular path of 5 m radius around the center of the square with the camera watching through the center of the circle; a first revolution provides 25 landscape images (1 m height above ground), while the second further 25 portrait images (1.9 m height above ground). The second camera follows two different circles, looking directly toward the buildings; while all images are landscape, the first 25 are acquired at 3.4 m height from the ground, and with a circle of radius 2 m, the last 25 follow a circle of radius 6.25 m at 4 m height. The third camera uses the same configuration as the first one, but the acquisition poses slightly differ in position and orientation. The fourth camera follows the border of the square taking pictures of its opposite side from 1.3 m above the ground. An overview of the paths followed by the cameras is visible in Fig. 8. These camera paths have been chosen to cover all of the facades and provide overlap between images acquired by the same camera as well as across different cameras. Cameras 1, 3, and 4 use a focal length of 35 mm, whilst Camera2 has a 50 mm focal length. The average GSD and depth for the cameras are, respectively: 8 mm @ 46 m, 5 mm @ 38 m, 8 mm @ 46 m, 10 mm @ 61 m. Different high dynamic range image (HRDI) maps were used for lighting the scene. Camera1 images are captured in a partly cloudy sky. Cameras 2 and 4 acquisitions use clear sky conditions. Camera3 images are acquired at sunrise, thus with a predominant orange color and strong shadows. Considering the availability of roughness and normals maps for different 3D models (in addition to the diffusive color component), we used Blender's Cycles path tracing engine to render photorealistic images of the scene.

The ENRICH-Statue dataset is based on a 3D model of a hunter created via Photogrammetry from 1016 photos and de-lighted with Agisoft Delighter. 4k albedo map, normal map, roughness map, concavity map, ambient occlusion map are also provided for texturing. The dataset uses the same virtual setup of ENRICH-Square, with an additional 2 m high 3D statue of a hunter placed at the center of the square. Cross pattern GCPs of size $2\ \text{cm} \times 2\ \text{cm}$ square are uniformly placed directly on the statue and its basement (Fig. 9). While their number is lower than the ones used in the ENRICH-Square dataset, each of them appears in at least 62 images. In this dataset, four cameras are used to acquire 200 pictures of the statue (50 images each). Some sample images are visible in Fig. 10a–b. Camera1 and Camera3 captured landscape images rotating around the statue (radius 3.75 m), looking at it slightly from the bottom (average height from ground 0.3 m). Camera2 and Camera4 rotated around the statue (radius 2.25 m), looking at it in portrait orientation from slightly above (height 1.9 m). The path followed by each camera is visible in Fig. 11. The path and orientations of the cameras ensure that the whole surface of the statue is covered in the images while also providing the overlap needed for the 3D reconstruction tasks. The average distance of the statue from the camera is 5.8 m for the first and third cameras, and 3.8 m for the second and fourth, thus providing a GSD on the statue of 0.69 mm, 0.70 mm, 0.64 mm, and 0.64 mm, respectively. As in the ENRICH-Square dataset, different HRDIs were used for lighting the scene. The whole scene is illuminated by a partly cloudy sky for images acquired by the first camera, and with sunrise lighting for those by the third camera. For Camera2 and Camera4 the light is provided by a sunny sky and a cloudy sky, respectively. Along with the RGB images, depth information is also generated for each picture taken. We used Blender's Cycles path tracing engine to render photorealistic images of the scene.

All the images have been rendered using an Nvidia Quadro RTX 6000 GPU. For each dataset, an additional depth information is obtained directly from the Rendering Layers of Blender (Figs. 4f, 7e, 10c).

3.2. Data description

Each ENRICH dataset includes rendered RGB images, depth files, the 3D model, and a set of text files.

The RGB images are available as JPEG files, named using the format <frame number>_<camera name>.jpg. The image files contain EXIF metadata regarding the camera model, its resolution, focal length, and camera pose.

The depth information corresponding to each RGB image is included as an OpenEXR file and a preview PNG file. The OpenEXR files provide

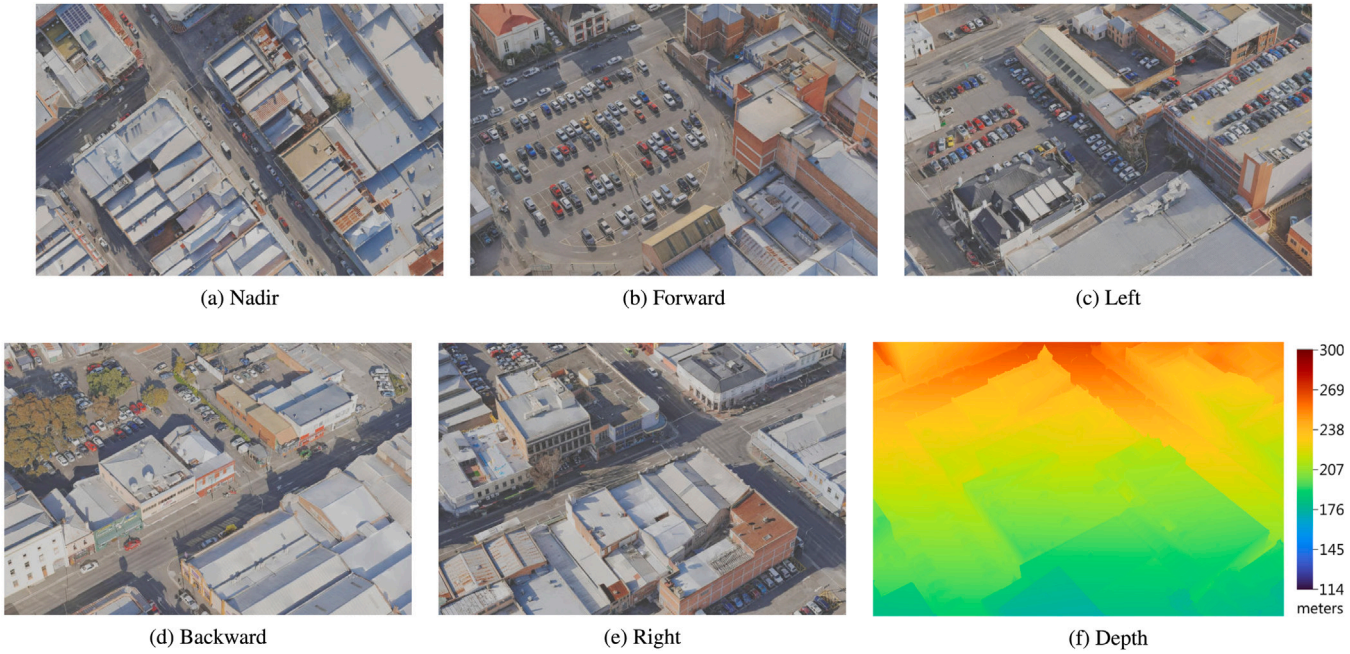


Fig. 4. Sample images from the ENRICH-Aerial dataset.

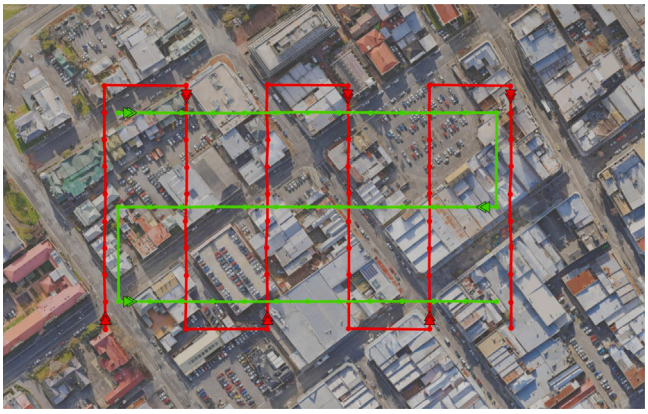


Fig. 5. Camera paths on the ENRICH-Aerial dataset. In red the path followed by the first nadir and oblique cameras, in green the path followed by the second nadir camera. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the ground truth metric distance (absolute depth) from the camera of each pixel in the image. Pixels depicting the sky use 10B as an infinity value. The PNG file provides a colored preview image of the depth using a different color at steps of 25 cm of distance from the camera position. The ENRICH-Aerial dataset uses a linear color map to show the depth variation. The ENRICH-Square and ENRICH-Statue datasets use a logarithmic color map to highlight the most relevant depth variations within a limited range of distances from the camera.

Black pixels are at infinity. Figs. 4f, 7e, and 10c show some examples of depth color images and their relative color mapping with respect to the distance from the camera in meters.

For each dataset, we provide two WavefrontOBJ files containing the 3D geometry of the scene. The first file contains only the geometry in the form of a 3D mesh of the whole scene (including the GCPs). The second one also includes material definitions and the related texture images in addition to the mesh.

Finally, some auxiliary text files supply information about the camera position in the 3D scene, GCPs, and their corresponding image coordinates. This information is provided in Comma and Tab-Separated-Values (CSV and TSV) files.

The `cameras.csv/.tsv` file describes the pose of the cameras using the following fields:

1. `label`: string, the filename of the image the entry refers to, including the extension.
2. `position_x,y,z`: three float numbers representing the global position of the camera.
3. `omega, phi, kappa`: Omega, Phi, Kappa angles defining the rotation of the camera. Three floats, representing angles in radians.
4. `yaw, pitch, roll`: rotation of the camera using Yaw, Pitch, Roll angles. Three floats, representing angles in radians.
5. `rotation_w,x,y,z`: four floats, representing the camera rotation as an Hamilton's quaternion.
6. `lookat_x,y,z`: three floats, representing the camera look-at direction vector.

The values are defined according to a global coordinate reference system, having X growing right/east, Y growing forward/north, and Z growing upward/zenith. Omega, Phi, Kappa are counterclockwise (CCW) local rotations along the X, Y, Z axis, applied in the following order $R = R_x \cdot R_y \cdot R_z$. Pitch and Roll are CCW local rotations, respectively, along the X and Y-axis, Yaw is a clockwise (CW) local rotation along the Z axis. The order of application is $R = R_z \cdot R_x \cdot R_y$. In any case, the provided parameters define the world-space rotation and translation. Thus, a camera with a 0 translation and a 0° rotation along all axes is placed at the origin and aligned to the global coordinate system (looking along the -Z axis with its up direction aligned with +Y and right direction aligned to +X). To map the 3D world from the camera-space is possible to use the equation $X_{cam} = R^T(X - X_0)$, where R is the rotation matrix of the camera, X is a 3D point in world coordinates, X_0 is the camera translation in world coordinates.

The `gcp.csv/.tsv` file describes the position of the GCPs in the scene. The reported fields are:

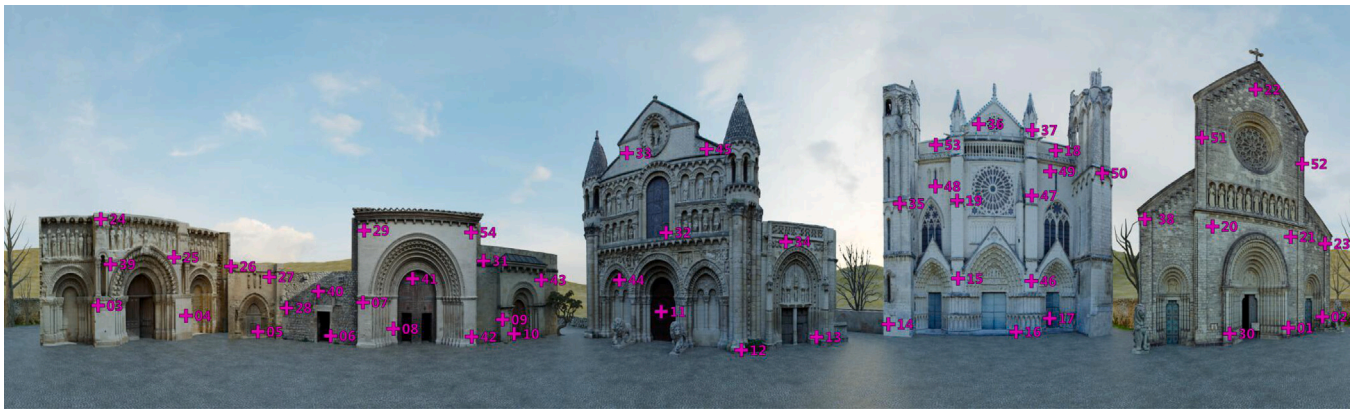


Fig. 6. Equirectangular projection of the ENRICH-Square scene showing GCPs placement.

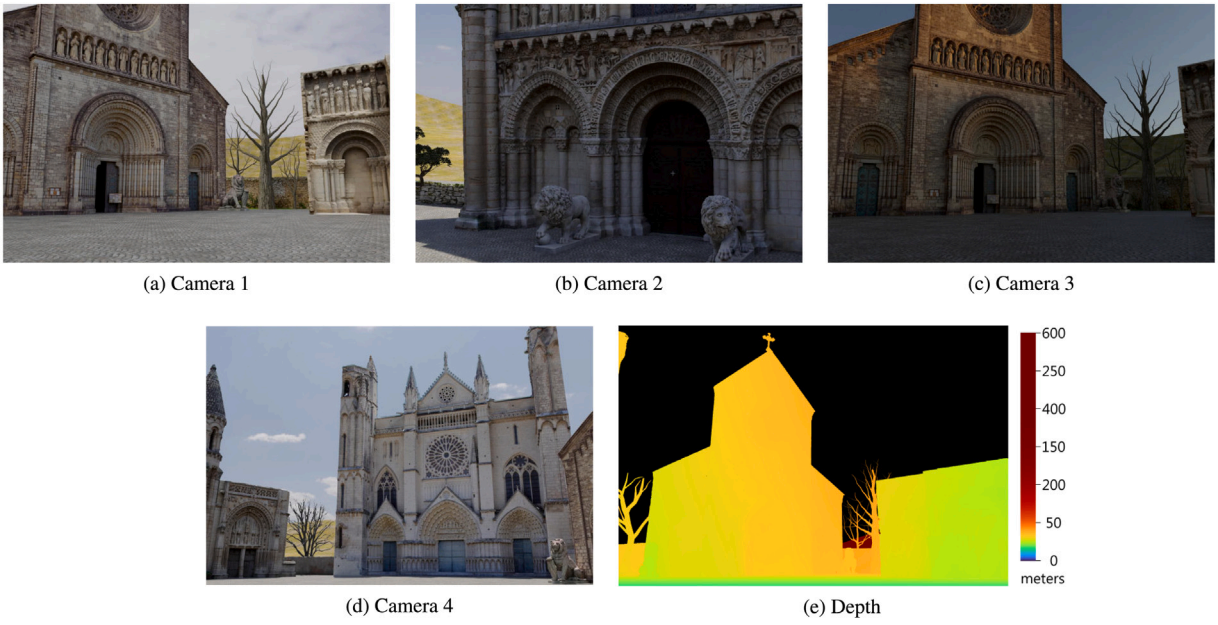


Fig. 7. Sample images from the ENRICH-Square dataset.

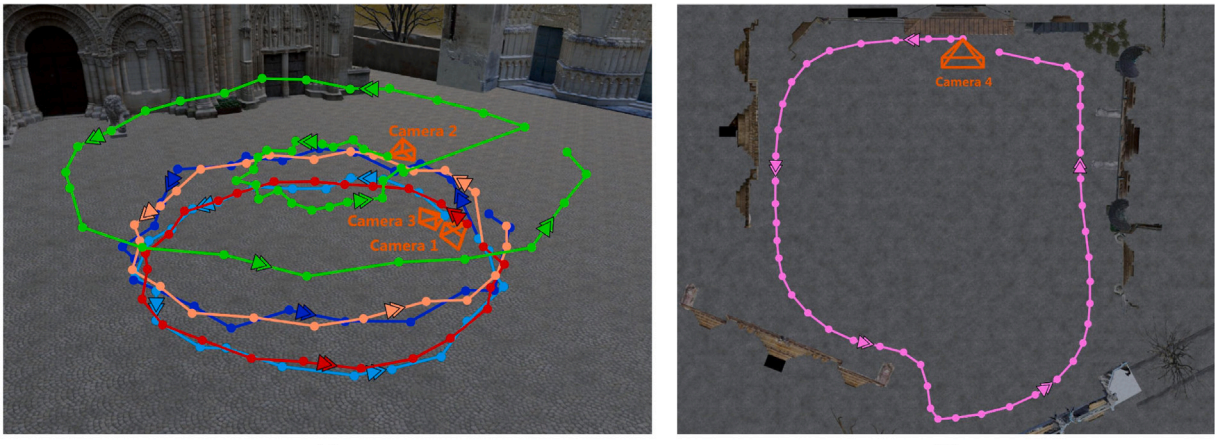


Fig. 8. Camera paths on the ENRICH-Square dataset. (a) Camera 1, 2, and 3. The path of cameras 1 and 3 change color where the camera rotates from landscape to portrait, red to orange and cyan to blue respectively. (b) Path of the 4th camera. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Fig. 9. Ground Control Point (GCP) locations on the ENRICH-Statue dataset.

(a) Camera 1 (b) Camera 2 (c) Depth

Fig. 10. Sample images from the ENRICH-Statue dataset.

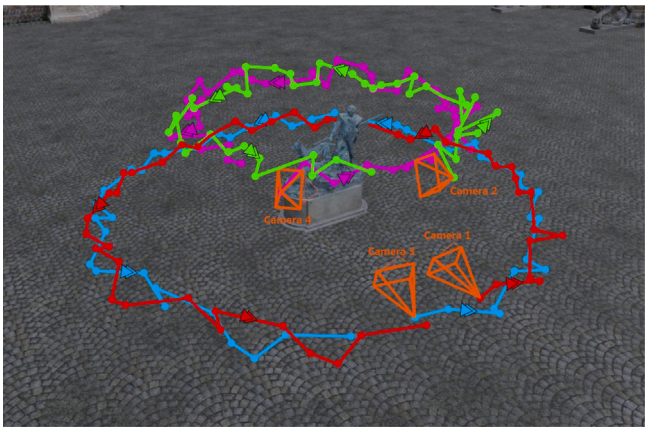


Fig. 11. Camera paths on the ENRICH-Statue dataset. The paths followed by cameras 1–4 are shown in red, green, cyan, and purple respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

1. `gcp_name`: string, unique identifier of the GCP.
2. `x_east, y_north, z_altitude`: three floats, representing the global coordinates of the center of the GCP in the scene.
3. `type`: string, defines the shape of the GCP either *cross* or *round*.

The `gcp_images_list.csv/.tsv` file reports the 2D coordinates of each GCP for each image in which is visible. The reported fields are:

1. `image_name`: string, the filename of the image that portrays the GCP.
2. `gcp_name`: string, unique identifier of the GCP.
3. `image_u,v`: two floats, u and v image coordinate of the center of the GCP in the images (0,0 is the top left corner of the top left pixel in the image, u grows right and v grows downwards).

4. Experiments

The ENRICH datasets enable researchers to test and compare algorithms for different photogrammetric and computer vision applications. The proposed and analyzed tasks reflect some recent interests of the research community, focused on exploring new solutions for image orientation, 3D scene reconstruction and quality assessment in large-scale mapping:

1. Evaluation of different local features, including both handcrafted and learning-based methods, for the estimation of camera poses (Section 4.1): the SfM pipeline is normally very robust, but a high overlap between images and constant lighting conditions are typically needed. New matching algorithms based on neural networks have been proposed to overcome these limitations and obtain robust matches even in very challenging conditions. However, further investigations on their behavior in different survey scenarios are still needed, considering their different performances with multi-temporal (Maiwald et al., 2021; Farella et al., 2022), wide-baseline (Bellavia et al., 2022b; Chen and Heipke, 2022), or aerial datasets (Remondino et al., 2022; Peppa et al., 2022). Our experiments focus on local features, but further SfM tasks have been recently revisited. As an example, new geometric verification approaches have been proposed in alternatives to RANSAC (Chum et al., 2005; Yi et al., 2018), or end-to-end deep learning-based methods which handle the entire SfM pipeline.
2. Analyses on the effect of the number and spatial distribution of Ground Control Points (GCPs) on 3D accuracy (Section 4.2): we investigate how GCP configurations can influence aerial triangulation (AT) and image orientation results, a topic not

fully explored. Especially in aerial mapping projects, the quality of the final products is highly correlated to GCPs number and distribution within the scene. Most of the available SfM tools estimate the exterior orientation parameters within a free-network bundle adjustment, followed by a 3D similarity transformation to move from an arbitrary to a real-world coordinate system. Investigations on the effects of GCPs spatial distribution on the 3D accuracy have been extensively proposed in the last years for UAV image blocks (Villanueva and Blanco, 2019; Garcia and Oliveira, 2020; Oniga et al., 2020; Ulvi, 2021) while a few studies have focused on their influence in the airborne case (Ostrowski and Bakuła, 2016; Gerke et al., 2016).

3. Monocular depth estimation for outdoor architectural scenarios (Section 4.3): the literature presents various methods for Monocular Depth Estimation (MDE) (Amiri et al., 2019; Tosi et al., 2019; Lee et al., 2019; Bhat et al., 2021; Welponer et al., 2022) for the prediction of depth maps from a single RGB image, without prior knowledge about the scene or camera parameters. The ENRICH dataset can represent an useful dataset for training and testing MDE algorithms. These methods are usually trained on task-specific datasets such as KITTI (Geiger et al., 2012) (autonomous driving) or NYUv2 (Silberman et al., 2012) (indoor environments). While training on such datasets is suitable for a given application, it limits the ability of the models to generalize the depth estimation on different scenes. Recently, some approaches (Li and Snavely, 2018; Ranftl et al., 2021) have improved generalization capabilities by using datasets depicting various scenes, from indoor environments to aerial views.

4.1. Evaluation of SfM pipelines

The SfM pipeline consists of numerous tasks (e.g. extraction of local features, image matching, and image orientation) that can be addressed by different algorithms. The ENRICH datasets offer the opportunity to assess their accuracy downstream of the reconstruction process instead of evaluating each task separately (Jin et al., 2021).

In this work, we investigate the efficacy of different local features within the SfM pipelines. In the last few years, new local features based on neural networks have appeared in addition to traditional methods. However, their performances, limits, and potentials for photogrammetric applications is still an open research topic. For example scenes characteristics (like texture types, illumination, scale, and camera rotation) are critical elements in conditioning the algorithm's performance. In our tests we consider and compare both deep learning-based features and traditional ones.

The images and ground truth of the ENRICH-Statue and ENRICH-Square datasets were used for the experiments. The two scenes represent typical terrestrial photogrammetric surveys, both in terms of surveyed objects and camera network. The available images are distortion-free and can be modeled as a pinhole camera with a known principal distance. Therefore, the evaluation relies on the accuracy of check points and external orientation parameters, neglecting considerations about the interior orientation. The check points are well-distributed targets on the scene whose 3D coordinates have been synthetically generated. In addition, the datasets offer images with relevant variations in scale, illumination, angle of view, and camera rotations, which are challenging situations for local features extractors.

For both datasets, we compared three different SfM pipelines:

1. `RootSIFT + COLMAP` (Schonberger and Frahm, 2016), an open-source SfM software, available both with a command line interface and a graphical user interface that allows the user to customize many SfM parameters. In this test, we used the build-in RootSIFT (Arandjelović and Zisserman, 2012) as local feature, followed by brute-force matching with near neighborhood ratio threshold set to 0.80, incremental reconstruction (resection–intersection), and local/global bundle adjustment.
2. `Deep learning local feature + COLMAP`. Several deep learning-based feature extractors have been tested, importing their keypoints and descriptors in COLMAP. Image matching and orientation have been performed with the default options, as in the previous method.
3. `Metashape`, the commercial software developed by Agisoft with its proprietary SfM pipeline implementation as reference.

For the comparison, 8000 local features were extracted on images resized to 1500×1000 pixels because of the high computational performance required by the deep learning-based methods. Tiling images is the alternative approach for dealing with full-size images, as proposed in Remondino et al. (2022).

When comparing different SfM pipelines, several metrics are possible. Often, the bundle statistics obtained downstream the SfM pipeline are used, (Schonberger et al., 2017), such as the number of correctly registered images, the number of triangulated 3D points, the mean track length (MTL) or multiplicity, the mean reprojection error (MRE), and the mean observations per image, which is the mean number of correct tie points. Previous works have highlighted how these metrics are often inconsistent with the actual accuracy of the reconstruction evaluated in the object space (Remondino et al., 2021; Bellavia et al., 2022a). Therefore, we evaluated the results in the object space, using the following metrics:

- root mean square error (RMSE) computed on a few well-recognizable object points (the targets provided by the ENRICH benchmark - check points, CPs),
- RMSE on the Centers of Projection (COP) of the cameras.

Another possible metric could be the RMSE on the camera angles. Note that currently COLMAP can perform only a bundle block adjustments in free network. For this reason, all targets have been used to compute only a Helmert transformation to get a scaled 3D result and not as constraints in the bundle adjustment solution. Therefore, in Sections 4.1.1 and 4.1.2 all the targets are used as CPs, and none as GCPs.

4.1.1. The ENRICH-Statue dataset

The ENRICH-Statue dataset presents images both in landscape and portrait orientation, requiring rotation invariant local features. As a deep learning approach, we chose LF-Net (Ono et al., 2018), an end-to-end convolutional neural network among the few local features trained to be invariant to rotations. Its results are compared against RootSIFT and Metashape.

In Table 3, the RMSE for the three pipelines are shown beside some bundle statistics reported for completeness. All the tested methods delivered a RMSE on the CPs of about one-third of the GSD (0.64–0.69 mm). As expected, worse results have been achieved for the 3D coordinates of the COPs, with a RMSE value similar to the GSD. Therefore, for this dataset, no significant differences among the three tested methods were found, apart from a slightly worse behavior of RootSIFT. In Fig. 12a is reported the camera network and the sparse point cloud processed by the RootSIFT implementation of COLMAP.

After the block orientation, the dense cloud was reconstructed for each method and compared in terms of cloud-to-cloud (C2C) distance with the point cloud extracted from the reference 3D models. Fig. 12b shows an example of a dense point cloud obtained with RootSIFT + COLMAP, with an achieved average Cloud-to-Cloud distance of 1.34 mm and a standard deviation of 0.74 mm. Comparable results in terms of C2C distance were obtained for LF-Net + COLMAP and Agisoft Metashape. This first test thus shows small differences in the performance of hand-crafted and deep learning-based local features both in terms of RMSE on check points and on the dense cloud accuracy.

Table 3
Bundle statistics and RMSE on CPs and COPs for different SfM pipelines using the ENRICH-Statue dataset.

Method	RMSE on CPs [mm]	RMSE on COPs [mm]	MTL	MRE on tie points [pix]	Total keypoints
RootSIFT + COLMAP	0.23	1.63	4.6	0.65	86,525
LF-Net + COLMAP	0.16	0.69	4.5	0.43	99,140
Metashape	0.14	0.89	3.6	0.65	77,994

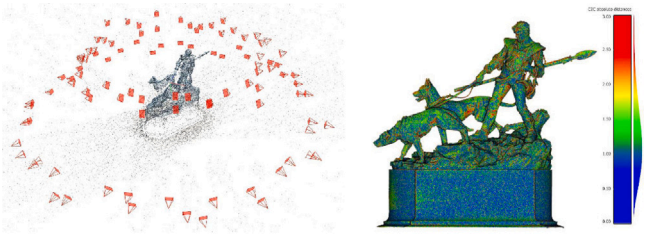


Fig. 12. Results obtained with RootSift + COLMAP. (a) Sparse cloud and camera network. (b) Dense cloud with C2C absolute distance in millimeters.

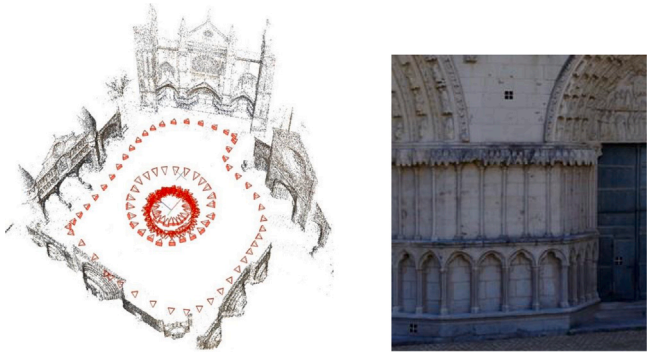


Fig. 13. Example of the ENRICH-Square sparse reconstruction performed with RootSIFT + COLMAP (a), and the detail of three white crosses on a black background used to materialize the targets (b).

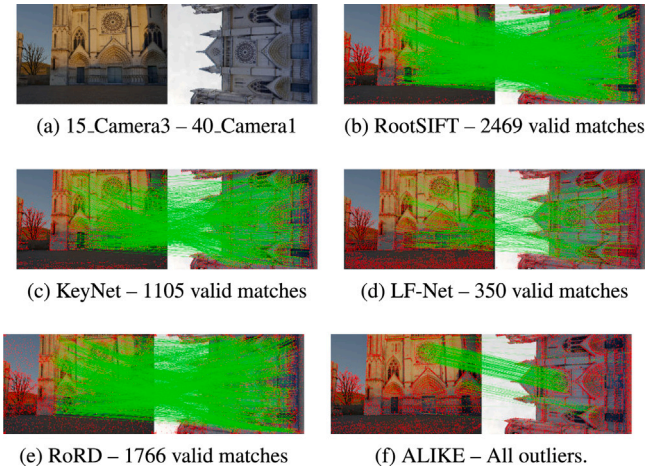


Fig. 14. Tie points extraction under rotation and illumination changes.

4.1.2. The ENRICH-Square dataset

In this section, we further deepen the local features comparison, using the ENRICH-Square dataset. Among the various deep learning-based methods, other rotational invariant features were chosen beside LF-Net: KeyNet + AffNet + HardNet (Barroso-Laguna et al., 2019; Mishkin et al., 2018; Pultar, 2020) and RoRD (Parihar et al., 2021). KeyNet, AffNet, and HardNet are respectively a deep learning-based detector, orientation estimator, and descriptor. In the rest of the paper we will refer to this combination simply as KeyNet. Rotation invariance is necessary to optimize the sensor calibration, although many deep learning-based methods still neglect this aspect. ALIKE (Zhao et al., 2022) is the most recent keypoint extractor, and it has been included in our tests, although it is not trained to be invariant to rotations.

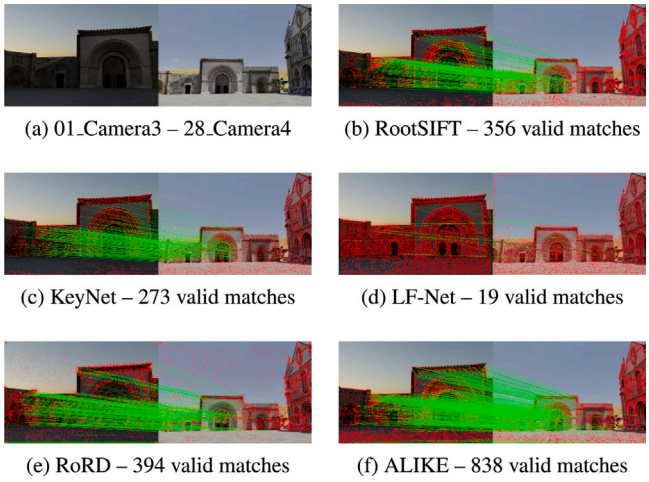


Fig. 15. Tie points extraction under scale and illumination changes.

Results are reported in Table 4 in terms of RMSE on CPs and COPs. In this test, RootSIFT still represents the state-of-the-art approach, with an RMSE of an order of magnitude lower than KeyNet, while all the other methods failed to register the entire image block due to relevant scale variation and sensor rotation. Agisoft Metashape performed similarly to RootSIFT. Fig. 13 shows the camera network and a closer view of a rectangular target (white cross on black background). As in Section 4.1.1, the accuracy on COPs is an order of magnitude larger than accuracy on CPs, and it does not seem to have the same sensitivity as that on CPs.

COLMAP provides further processing statistics, also reported in Table 4. It is worth noting the excessively high mean reprojection error (MRE) of RoRD, equal to 1.298 pixels, and how KeyNet has the highest mean track length (MTL), even if it extracted fewer keypoints (6775 kpts per image, less than the required 8000).

Figs. 14, 15, 16, and 17 show a few image pairs representing interesting challenging conditions for the local features.

Fig. 14 shows the case of a 90 degrees rotation and lighting changes. Except for ALIKE, all others methods handled sensor rotations properly, showing that ALIKE is not trained for rotation invariance. Note how the number of LF-Net valid matches is significantly lower than the other methods. Fig. 15 shows a significant scale variation and changes in illumination, adequately addressed by all the methods with the LF-Net exception. In Fig. 15e can also be noted that RoRD is the only method to find keypoints in homogeneous sky areas. Fig. 16 remarkably accentuates the scaling factor with respect to Fig. 15, demonstrating that the scale is a further critical factor for learning-based methods, in addition to the rotation invariance (Remondino et al., 2021; Bellavia et al., 2022b). Finally, Fig. 17 reports the case of an extreme perspective

Table 4
RMSE and statistics for the comparison of deep learning-based and hand-crafted local features with the ENRICH-Square dataset.

Method	RMSE on CPs [cm]	RMSE on COPs [cm]	MTL	MRE on tie points [pix]	Total keypoints
RootSIFT + COLMAP	0.333	2.597	8.47	0.306	1,597,879
KeyNet + COLMAP	1.523	2.891	9.37	0.680	1,354,898
LF-Net + COLMAP	Failed to register all images		5.59	0.647	1,600,000
RoRD + COLMAP	Failed to register all images		6.37	1.298	1,454,954
ALIKE + COLMAP	Failed to register all images		6.81	0.689	1,600,000
Metashape	0.286	1.982	4.66	0.428	1,600,000

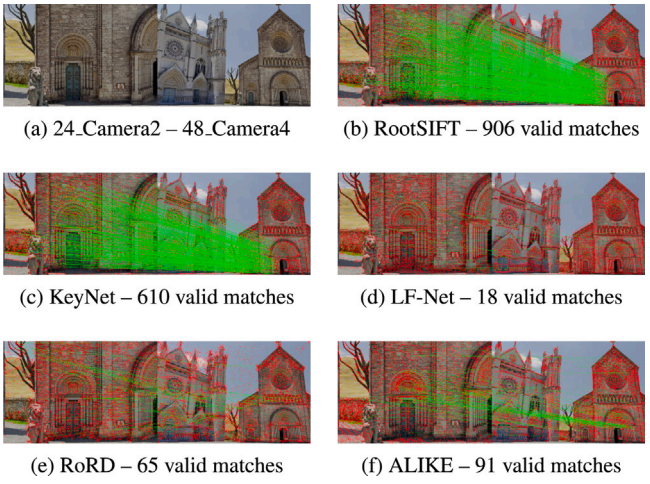


Fig. 16. Tie points extraction under large scale changes.

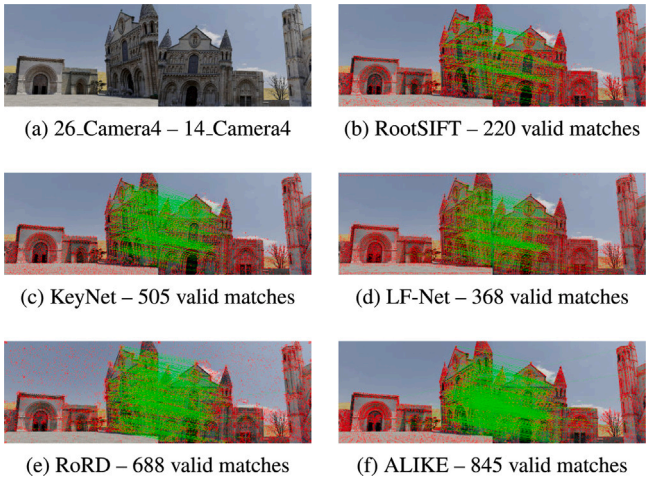


Fig. 17. Tie points extraction under large view changes.

variation and shows the good performance of deep learning-based methods in managing these situations.

This section showed how the ENRICH dataset provides a variety of challenging conditions useful for testing new SfM algorithms, such as local detectors and descriptors. In particular, it was shown that the new deep learning-based local features sometimes fail to orient the whole image block or are less accurate than classical methods such as RootSIFT, because of strong rotations or scale variations.

4.2. The effects of Ground Control Points spatial distribution on the 3D accuracy

This section examines the impact of Ground Control Points (GCPs) number and spatial distribution within the scene on 3D accuracy, exploiting the ENRICH-Aerial dataset.

In aerial triangulation (AT), GCPs are used to georeference data and optimize camera orientation by providing additional information for refining the bundle adjustment. Their configuration (number and distribution) affects the reconstruction results. Most of the recent and available automatic processing solutions for exterior orientation initially adopt a free-network approach for the bundle adjustment, and GCP coordinates are then introduced for georeferencing data and refining the orientation results. This approach was followed in our tests for the exterior orientation, while no further considerations were made for the interior orientation, the images being without distortion.

Different GCPs configurations were investigated for processing the block of 300 oblique and nadir images at the original image size (6016× 4016 pixels). In our experiments, a variable and increasing number of targets (four to twelve) with different distributions were used as GCPs, while the rest as Check Points (CPs). Both the image and spatial coordinates provided in the dataset were employed. Tested configurations and distribution schemes (Fig. 18) are defined as follows:

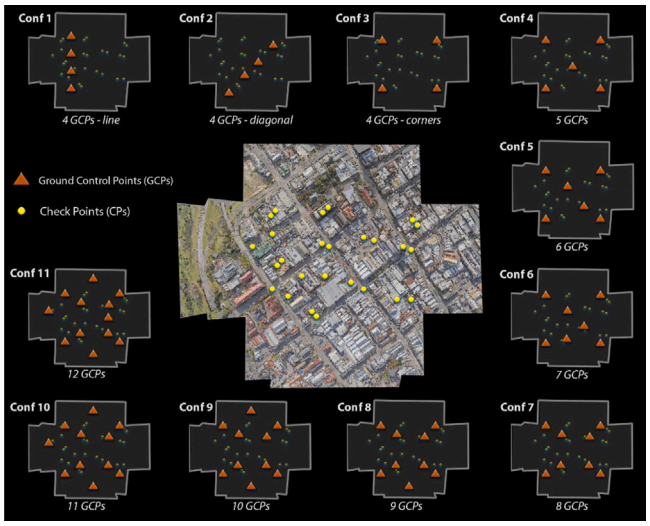


Fig. 18. Schemes of the eleven GCPs configurations, varying in the number and distribution of ground and check points.

- Configuration 1: four GCPs aligned on the shorter side of the block;
- Configuration 2: four GCPs aligned along the diagonal of the block;
- Configuration 3: four GCPs distributed on the edges of the block;
- Configurations 4 to 11: an increasing number of GCPs (from five to twelve), distributed on the edges, and progressively adding points inside and outside the central area of the block.

In order to assess the achieved accuracy for each configuration, the root mean square error (RMSE) between the ground truth and the

Table 5
RMSE of the planimetric (Rx and Ry), vertical (Rz) and global (Rx,y,z) residuals on Check Points (CPs) with the eleven Ground Control Points (GCPs) configurations.

	Rx [mm]	Ry [mm]	Rz [mm]	Rxyz [mm]
Conf 1 (4 GCPs line)	15.085	5.849	19.073	25.011
Conf 2 (4 GCPs diagonal)	3.436	5.782	20.195	21.286
Conf 3 (4 GCPs edges)	3.812	3.372	4.012	6.481
Conf 4 (5 GCPs)	4.028	3.509	3.340	6.300
Conf 5 (6 GCPs)	3.725	3.470	3.351	6.095
Conf 6 (7 GCPs)	3.730	3.553	3.452	6.201
Conf 7 (8 GCPs)	3.592	3.606	3.494	6.174
Conf 8 (9 GCPs)	3.240	3.627	3.417	5.944
Conf 9 (10 GCPs)	2.853	3.036	3.155	5.227
Conf 10 (11 GCPs)	3.067	2.967	2.722	5.062
Conf 11 (12 GCPs)	1.931	3.075	2.641	4.490

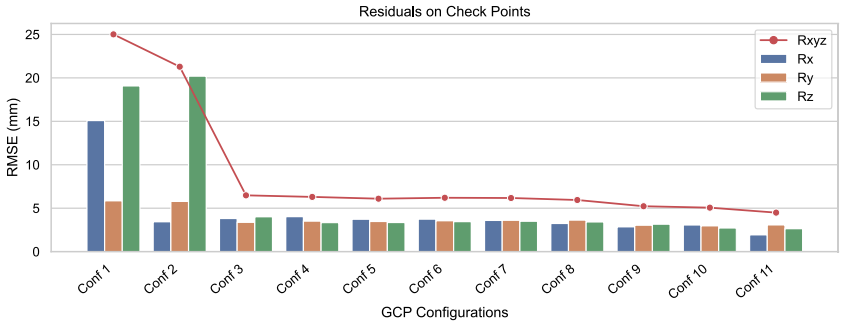


Fig. 19. RMSE residuals on Check Points (CPs) with eleven GCPs configurations.

estimated 3D coordinates after the bundle adjustment was calculated for the CPs. RMSE of planimetric (Rx and Ry), vertical (Rz), and 3D (Rxyz) errors are reported in Table 5 and visualized in Fig. 19.

As expected, the worst results on the CPs occurred when four GCPs were arranged along one line (Configurations 1 and 2), with a significant improvement of metrics when the same number is uniformly displaced on the edges of the block (Configuration 3). Increasing the number of GCPs and their spread distribution within the area generates less relevant changes in error metrics. While Configuration 4 confirms that adding a point in the middle of the block increases the altimetric accuracy, a more marked improvement is visible only in the last configuration. Furthermore, the eleven sparse point clouds were compared with the reference mesh model provided in the ENRICH-Aerial dataset for a more in-depth analysis and quality assessment of the results achieved with the different GCPs configurations. The orthogonal distance between each point and the corresponding triangle surface for each configuration scenario returned no significant differences in the standard deviation (for all the cases, around 0.1 m). In contrast, more relevant divergences are evident when comparing the average distances, as shown in Fig. 20. Also for this case, the worst metrics are related to the first two configurations, with a clear improvement in the other scenarios. Tests with the ENRICH-Aerial dataset show that 3D accuracy is more affected by the GCPs spatial distribution with respect to their number. Four GCPs distributed on the edges of the aerial block are already sufficient for a clear improvement of the error metrics.

4.3. Monocular depth estimation

Monocular Depth Estimation (MDE) is the task of estimating the distance of the surface depicted by each pixel in a single RGB image. This task is of interest for many fields, most notably 3D scene reconstruction, autonomous driving, and Augmented Reality. Among the multitude of MDE methods, we selected MegaDepth (Li and Snavely, 2018) and Dense Prediction Transformers (DPT) (Ranftl et al., 2021). MegaDepth proposes a large depth dataset built by using structure-from-motion and multi-view stereo on internet photo collections, seeking to learn to predict monocular depth with high accuracy and generalizability. It depicts different city landmarks, including buildings, statues, and squares. In addition to the dataset, the authors train different architectures obtaining the best depth estimation results with the hourglass architecture proposed by Chen et al. (2016). DPT proposes the use of dense vision transformers for the prediction of monocular depth, showing an improvement of up to 28% in relative performance when compared to a state-of-the-art fully-convolutional network. In addition to the architecture, the authors also propose the use of a meta-dataset to train the neural network. This dataset is composed of 10 different depth datasets, each including different scene types ranging from indoor to outdoor and even aerial views.

We evaluate MegaDepth and DPT on the three ENRICH datasets, we used all of the available images of each dataset. We follow the procedure defined by Ranftl et al. (2020) to evaluate the results of both methods. The images of the dataset are resized to match the input size required by each approach. The prediction is resized to the ground truth resolution using nearest-neighbor interpolation before the evaluation. Pixels belonging to the sky in the ground truth are used as a mask to exclude invalid portions from the prediction as well as the ground truth during evaluation. Since the prediction and the ground truth may differ in scale and shift, we align the predictions before measuring errors. We perform this alignment in the inverse-depth space based on the least-squares criterion as in Ranftl et al. (2020). Scale s and shift t factors are determined as $(s,t) = \text{argmin}_{s,t} \sum_{i=1}^{N} (sd_i + t - d_i^*)$, where N denotes the number of pixels, d is the predicted relative inverse-depth, and d^* is the ground truth inverse-depth. We provide evaluation results using a depth cap suitable for each dataset as well as non-capped predictions. Such depth cap is used to include over far pixels in the exclusion mask.

The first metric used for the evaluation is the mean absolute value of the relative error $AbsRel = (1/M) \sum_{i=1}^{M} |z_i - z_i^*|/z_i^*$ in depth space. M denotes the number of valid pixels (not masked), z is the aligned predicted relative depth, and z^* is the ground truth absolute depth. The second metric is the percentage of pixels with $\delta = max\left(\frac{z_i}{z_i^*}, \frac{z_i^*}{z_i}\right) > \theta$ in depth space. θ defines a threshold for the evaluation and a common value is θ=1.25, which considers wrong pixels only those whose difference in depth is more than 25% of the ground truth value.

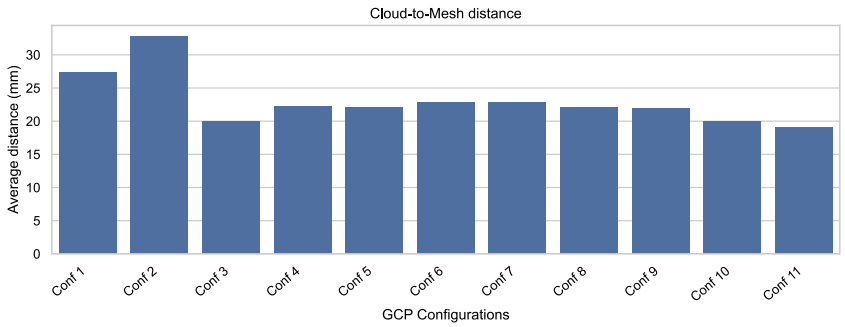


Fig. 20. Average Cloud-to-Mesh distance (mm) between the eleven sparse point clouds computed in different GCPs configuration scenarios and the reference mesh model provided in the ENRICH-Aerial dataset.

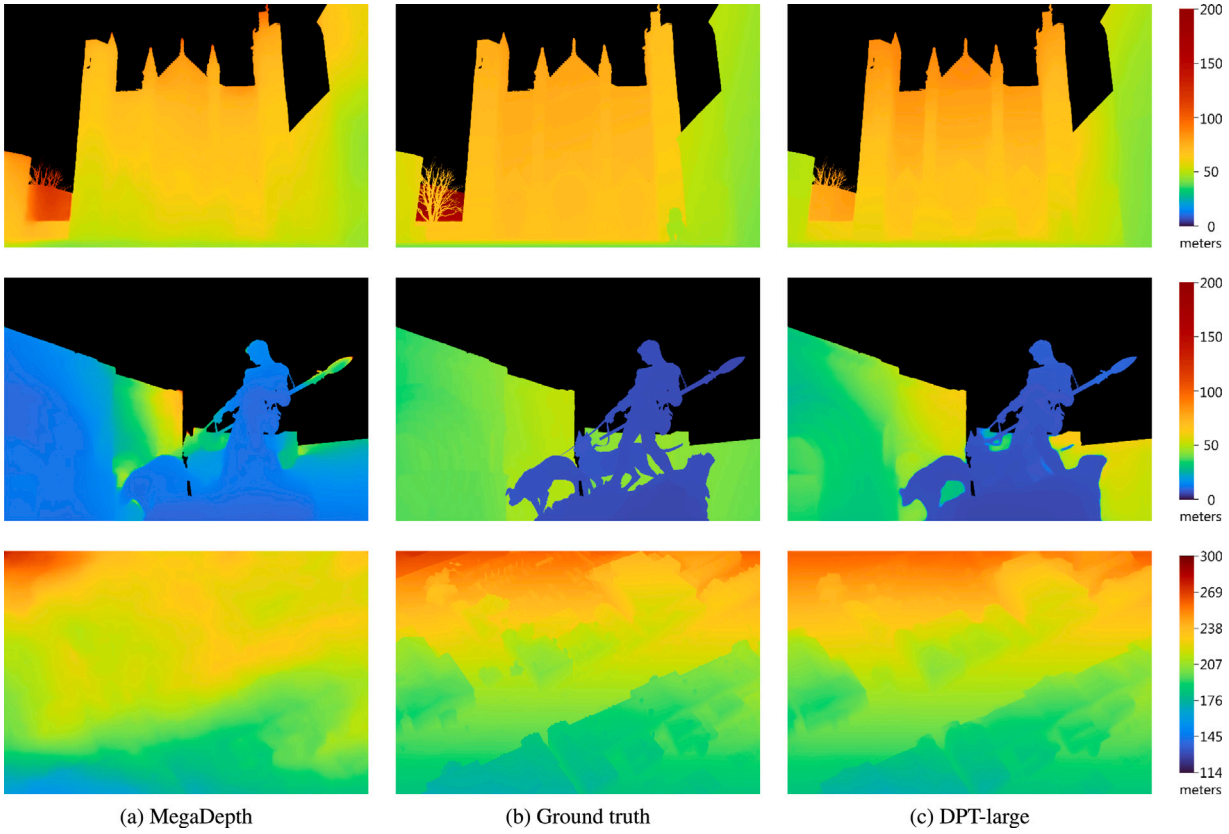


Fig. 21. Example of predicted depth maps for the ENRICH-Square, ENRICH-Statue, and ENRICH-Aerial datasets.

For the experiments, we used the codes and the pre-trained models provided by the authors. For MegaDepth and DPT, we used the best-generalization and the DPT-large models respectively. Examples of depth predictions as well as ground truth are visible in Fig. 21. Tables 6 and 7 report evaluation results on the ENRICH-Statue and ENRICH-Square respectively. For both datasets, we evaluate the predictions with a 70 m depth cap and with uncapped depths. To evaluate the effect of image orientation on depth estimation, we tested the two methods with portrait images either rotated or not. For both datasets and in all the experiment configurations, the best results are obtained by DPT. Table 8 reports the results of evaluation on the ENRICH-Aerial dataset. On this dataset, we performed the evaluation with uncapped depth ground truth. Since δ>1.25 allows an error up to 37 m for the nadir cameras and 54 m for the oblique cameras at the average depths, we also report the δ>1.05 value, which allows an error up to 7 m for the nadir cameras and 10 m for the oblique cameras.

DPT-large achieved lower errors than MegaDepth on all three datasets. Both methods achieve the best results on the ENRICH-Square dataset. This is mainly related to the scene setup that resembles some of the data used for the training of the Neural Networks. The ENRICH-Statue dataset is challenging for both methods: while MegaDepth fails to estimate the correct relative depth order of the elements in the scene, DPT-large has difficulties in correctly identifying and separating the foreground statue from the background elements. The evaluation on the ENRICH-Aerial dataset shows low errors for the δ>1.25, but this error increase significantly when the threshold is reduced to 1.05. Even in this case, DPT-large provides depth estimation with a lower error, and the result of MegaDepth on the aerial view appears flat and fails to highlight the buildings from the ground. Finally, while the depth

Table 6
Evaluation of the depth estimation on ENRICH-Square.

Method	Portrait as landscape				Portrait as portrait			
	Depth cap 70 m		No depth cap		Depth cap 70 m		No depth cap	
	AbsRel	$\delta > 1.25$	AbsRel	$\delta > 1.25$	AbsRel	$\delta > 1.25$	AbsRel	$\delta > 1.25$
MegaDepth	0.111	13.986%	0.211	16.918%	0.108	13.041%	0.208	16.290%
DPT-large	0.087	10.337%	0.135	13.369%	0.085	9.869%	0.134	12.975%

Table 7
Evaluation of the depth estimation on ENRICH-Statue.

Method	Portrait as landscape				Portrait as portrait			
	Depth cap 70 m		No depth cap		Depth cap 70 m		No depth cap	
	AbsRel	$\delta > 1.25$	AbsRel	$\delta > 1.25$	AbsRel	$\delta > 1.25$	AbsRel	$\delta > 1.25$
MegaDepth	0.577	86.278%	1.217	86.530%	0.459	71.832%	1.227	72.055%
DPT-large	0.244	40.358%	0.287	41.635%	0.146	20.133%	0.188	21.155%

Table 8
Evaluation of the depth estimation on ENRICH-Aerial.

Method	AbsRel	$\delta > 1.25$	$\delta > 1.05$
MegaDepth	0.039	0.060%	27.708%
DPT-large	0.017	0.001%	4.824%

cap of the predictions has a limited impact on the evaluation, the correct rotation of the portrait images significantly influences the depth estimation and evaluation. Providing the ENRICH-Statue images in the wrong orientation notably affects the results of DPT-large.

5. Conclusions

In this work we presented the ENRICH dataset, a multi-purpose set of synthetic images realized to complement existing terrestrial and aerial datasets. It provides challenging data to boost several research activities in photogrammetry and computer vision fields. ENRICH comprises three sets of data featuring images with different formats, cameras, environmental and acquisition conditions. Besides the rendered images, ENRICH includes also GCP coordinates, depth maps, and 3D models as pixel-precise ground truth.

The variety of data provided by ENRICH is suitable for testing methods and algorithms designed for different application domains, such as remote sensing, photogrammetry, and computer vision. Our experiments show that it can be effectively used in several challenging tasks, including SfM, MVS, MDE, etc. One uniqueness of ENRICH is the availability of multi-scale data, from aerial to terrestrial, enabling the evaluation and comparison of methods and algorithms under various conditions and gathering insights on their strengths and limits.

We plan to extend ENRICH further to include monocular and stereo video sequences for the two terrestrial scenes, semantic segmentation information, and camera distortions to allow the development and benchmarking of, for example, SLAM and 3D classification methods.

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

This research was supported by grants from NVIDIA and utilized NVIDIA Quadro RTX 6000. This work was also partly supported by the project "AI@TN" funded by the Autonomous Province of Trento (Italy). Authors are thankful to Michele Welponer (3DOM-FBK) for contributing in the preparation of an initial 3D scene the further elaborated and included in ENRICH.

References

Aanæs, H., Dahl, A.L., Steenstrup Pedersen, K., 2012. Interesting interest points. Int. J. Comput. Vis. 97 (1), 18–35.

Amiri, A.J., Loo, S.Y., Zhang, H., 2019. Semi-supervised monocular depth estimation with left-right consistency using deep neural network. In: 2019 IEEE International Conference on Robotics and Biomimetics. ROBIO, IEEE, pp. 602–607.

Arandjelović, R., Zisserman, A., 2012. Three things everyone should know to improve object retrieval. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 2911–2918.

Bakula, K., Mills, J., Remondino, F., 2019. A review of benchmarking in photogrammetry and remote sensing. Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.

Barroso-Laguna, A., Riba, E., Ponsa, D., Mikolajczyk, K., 2019. Key. net: Keypoint detection by handcrafted and learned cnn filters. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5836–5844.

Bellavia, F., Colombo, C., Morelli, L., Remondino, F., 2022a. Challenges in image matching for cultural heritage: an overview and perspective. In: roceedings of the 2nd International Workshop on Fine Art Pattern Extraction and Recognition (FAPER2022), Accepted.

Bellavia, F., Morelli, L., Menna, F., Remondino, F., 2022b. Image orientation with a hybrid pipeline robust to rotations and wide-baselines. Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. 46, 73–80.

Bhat, S.F., Alhashim, I., Wonka, P., 2021. Adabins: Depth estimation using adaptive bins. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4009–4018.

Bianco, S., Ciocca, G., Marelli, D., 2018. Evaluating the performance of structure from motion pipelines. J. Imaging 4 (8).

Blender Online Community, 2018. Blender - a 3D modelling and rendering package. Blender Foundation, Stichting Blender Foundation, Buikslotermeerplein 161, Amsterdam, Netherlands, URL: http://www.blender.org.

Blender Online Community, 2021. Eevee - Blender Manual. Blender Foundation, URL: https://docs.blender.org/manual/en/2.93/render/eevee/index.html.

Chen, W., Fu, Z., Yang, D., Deng, J., 2016. Single-image depth perception in the wild. Adv. Neural Inf. Process. Syst. 29.

Chen, L., Heipke, C., 2022. Deep learning feature representation for image matching under large viewpoint and viewing direction change. ISPRS J. Photogramm. Remote Sens. 190, 94–112.

Chum, O., Werner, T., Matas, J., 2005. Two-view geometry estimation unaffected by a dominant plane. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). 1, IEEE, pp. 772–779.

Farella, E., Morelli, L., Remondino, F., Mills, J., Haala, N., Crompvoets, J., 2022. The EUROSDR time benchmark for historical aerial images. Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. 43, 1175–1182.

Garcia, M., Oliveira, H., 2020. the influence of ground control points configuration and camera calibration for DTM and orthomosaic generation using imagery obtained from a low-cost UAV. ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci. 5 (1).

Geiger, A., Lenz, P., Urtasun, R., 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 3354–3361.

Gerke, M., Nex, F., Remondino, F., Jacobsen, K., Kremer, J., Karel, W., Huf, H., Ostrowski, W., 2016. Orientation of oblique airborne image sets-experiences from the ISPRS/EUROSDR benchmark on multi-platform photogrammetry. Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. 41-B1 41, 185–191.

Jensen, R., Dahl, A., Vogiatzis, G., Tola, E., Aanæs, H., 2014. Large scale multi-view stereopsis evaluation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 406–413.

Jin, Y., Mishkin, D., Mishchuk, A., Matas, J., Fua, P., Yi, K.M., Trulls, E., 2021. Image matching across wide baselines: From paper to practice. Int. J. Comput. Vis. 129 (2), 517–547.

Knapitsch, A., Park, J., Zhou, Q.-Y., Koltun, V., 2017. Tanks and temples: Benchmarking large-scale scene reconstruction. ACM Trans. Graph. (ToG) 36 (4), 1–13.

Launceston City Council, 2017. Central Launceston photo mesh 3D Model. https://data.gov.au/data/dataset/71e3b134-5fa2-466f-8fc9-87f900e87639 (Accessed on 13 May 2022).

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444.

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.

Li, Z., Snavely, N., 2018. MegaDepth: Learning single-view depth prediction from internet photos. In: Computer Vision and Pattern Recognition. CVPR, pp. 2041–2050.

Maiwald, F., Lehmann, C., Lazariv, T., 2021. Fully automated pose estimation of historical images in the context of 4D geographic information systems utilizing machine learning methods. ISPRS Int. J. Geo-Inf. 10 (11), 748.

Marelli, D., Bianco, S., Ciocca, G., 2020. IVL-SYNTHSFM-v2: A synthetic dataset with exact ground truth for the evaluation of 3D reconstruction pipelines. Data in Brief 105041.

Marelli, D., Bianco, S., Ciocca, G., 2022. SfM Flow: A comprehensive toolset for the evaluation of 3D reconstruction pipelines. SoftwareX 17, 100931.

Mishkin, D., Radenovic, F., Matas, J., 2018. Repeatability is not enough: Learning affine regions via discriminability. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 284–300.

Nikolenko, S.I., et al., 2021. Synthetic Data for Deep Learning. Springer.

Oniga, V.-E., Breaban, A.-I., Pfeifer, N., Chirila, C., 2020. Determining the suitable number of ground control points for UAS images georeferencing by varying number and spatial distribution. Remote Sens. 12 (5), 876.

Ono, Y., Trulls, E., Fua, P., Yi, K.M., 2018. LF-Net: Learning local features from images. Adv. Neural Inf. Process. Syst. 31.

Ostrowski, W., Bakuła, K., 2016. Towards efficiency of oblique images orientation. Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. XL-3/W4, 91–96.

Özdemir, E., Toschi, I., Remondino, F., 2019. A multi-purpose benchmark for photogrammetric urban 3D reconstruction in a controlled environment. In: Evaluation and Benchmarking Sensors, Systems and Geospatial Data in Photogrammetry and Remote Sensing. 42, pp. 53–60.

Parihar, U.S., Gujarathi, A., Mehta, K., Tourani, S., Garg, S., Milford, M., Krishna, K.M., 2021. RoRD: Rotation-robust descriptors and orthographic views for local feature matching. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems. IROS, IEEE, pp. 1593–1600.

Peppa, M., Morelli, L., Mills, J., Penna, N., Remondino, F., 2022. Handcrafted and learning-based tie point features–comparison using the EUROSDR RPAS benchmark dataset. Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. 43, 1183–1190.

Pultar, M., 2020. Improving the HardNet descriptor. arXiv preprint arXiv:2007.09699.

Ranftl, R., Bochkovskiy, A., Koltun, V., 2021. Vision transformers for dense prediction. ArXiv Preprint.

Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V., 2020. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE Trans. Pattern Anal. Mach. Intell..

Remondino, F., Menna, F., Morelli, L., 2021. Evaluating hand-crafted and learning-based features for photogrammetric applications. Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. 43, 549–556.

Remondino, F., Morelli, L., Stathopoulou, E., Elhashash, M., Qin, R., 2022. AErial triangulation with learning-based tie points. Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. 43, 77–84.

Rupnik, E., Nex, F., Toschi, I., Remondino, F., 2015. Aerial multi-camera systems: Accuracy and block triangulation issues. ISPRS J. Photogramm. Remote Sens. 101, 233–246.

Scharstein, D., Hirschmüller, H., Kitajima, Y., Krathwohl, G., Nešić, N., Wang, X., Westling, P., 2014. High-resolution stereo datasets with subpixel-accurate ground truth. In: German Conference on Pattern Recognition. Springer, pp. 31–42.

Schonberger, J.L., Frahm, J.-M., 2016. Structure-from-motion revisited. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4104–4113.

Schonberger, J.L., Hardmeier, H., Sattler, T., Pollefeys, M., 2017. Comparative evaluation of hand-crafted and learned local features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1482–1491.

Schops, T., Schonberger, 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. pp. 3260–3269.

Seitz, S.M., 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 (CVPR'06). 1, IEEE, pp. 519–528.

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.

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. Ieee, pp. 1–8.

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. pp. 9799–9809.

Tremblay, J., Prakash, A., Acuna, D., Brophy, M., Jampani, V., Anil, C., To, T., Cameracci, E., Boochoon, S., Birchfield, S., 2018. Training deep networks with synthetic data: Bridging the reality gap by domain randomization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 969–977.

Ulvi, A., 2021. The effect of the distribution and numbers of ground control points on the precision of producing orthophoto maps with an unmanned aerial vehicle. J. Asian Archit. Build. Eng. 20 (6), 806–817.

Vasiljevic, I., Kolkin, N., Zhang, S., Luo, R., Wang, H., Dai, F.Z., Daniele, A.F., Mostajabi, M., Basart, S., Walter, M.R., Shakhnarovich, G., 2019. DIODE: A dense indoor and outdoor DEpth dataset. CoRR arXiv:1908.00463.

Villanueva, J., Blanco, A., 2019. Optimization of ground control point (GCP) configuration for unmanned aerial vehicle (UAV) survey using structure from motion (SFM). Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. 42, 167–174.

Welponer, M., Stathopoulou, E., Remondino, F., 2022. Monocular depth prediction in photogrammetric applications. Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. 43, 469–476.

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: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1790–1799.

Yi, K.M., Trulls, E., Ono, Y., Lepetit, V., Salzmann, M., Fua, P., 2018. Learning to find good correspondences. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2666–2674.

Zhao, X., Wu, X., Miao, J., Chen, W., Chen, P.C., Li, Z., 2022. Alike: Accurate and lightweight keypoint detection and descriptor extraction. IEEE Trans. Multimed.