Endomapper dataset of complete calibrated endoscopy procedures

Pablo Azagra[1,3], Carlos Sostres[2,3,4,5], Angel Ferrandez[2,3,4,5], Luis Riazuelo[1,3], Clara Tomasini[1,3], Óscar León Barbed[1,3], Javier Morlana[1,3], David Recasens[1,3], Víctor M. Batlle[1,3], Juan J. Gómez-Rodríguez[1,3], Richard Elvira[1,3], Julia Lopez[2,3], Cristina Oriol[1,3], Javier Civera[1,3], Juan D. Tardós[1,3], Ana C. Murillo[1,3], Angel Lanas[2,3,4,5], and Jose M.M. Montiel[1,3]

[1]Instituto de Investigación en Ingeniería de Aragón (I3A)
[2]Hospital Clínico Universitario Lozano Blesa, Zaragoza, Spain
[3]University of Zaragoza, Spain
[4]IIS Aragon
[5]Ciberhd

ABSTRACT

Computer-assisted systems are becoming broadly used in medicine. In endoscopy, most research focuses on automatic detection of polyps or other pathologies, but localization and navigation of the endoscope is completely performed manually by physicians. To broaden this research and bring spatial Artificial Intelligence to endoscopies, data from complete procedures are needed. This data will be used to build a 3D mapping and localization systems that can perform special task like, for example, detect blind zones during exploration, provide automatic polyp measurements, guide doctors to a polyp found in a previous exploration and retrieve previous images of the same area aligning them for easy comparison. These systems will provide an improvement in the quality and precision of the procedures while lowering the burden on the physicians.
This paper introduces the Endomapper dataset, the first collection of complete endoscopy sequences acquired during regular medical practice, including slow and careful screening explorations, making secondary use of medical data. Its original purpose is to facilitate the development and evaluation of VSLAM (Visual Simultaneous Localization and Mapping) methods in real endoscopy data. The first release of the dataset is composed of 59 sequences with more than 15 hours of video. It is also the first endoscopic dataset that includes both the computed geometric and photometric endoscope calibration with the original calibration videos. Meta-data and annotations associated to the dataset varies from anatomical landmark and description of the procedure labeling, tools segmentation masks, COLMAP 3D reconstructions, simulated sequences with groundtruth and meta-data related to special cases, such as sequences from the same patient. This information will improve the research in endoscopic VSLAM, as well as other research lines, and create new research lines.

1 Background & Summary

Endoscopes traversing body cavities are routine. However, their potential for navigation assistance or device autonomy remains mostly locked. A computer-assisted endoscope requires spatial artificial intelligence, i.e. a map of the regions where it is navigating, along with the endoscope localization within this map. This capability is known in robotics literature as VSLAM (Simultaneous Localization and Mapping from Visual sensors). VSLAM will provide endoscopy with augmented reality, detection of blind zones, polyp measurements or guidance to a polyp found in previous explorations. In the long term, VSLAM will support utterly new robotized autonomous procedures.

VSLAM goal is building a per-patient map, in real-time, during the insertion of the first procedure. This map will be exploited and perfected either during the withdrawal of the first procedure or in any other future procedure.

There are mature methods for out of the body VSLAM[1,2]. However, bringing them to endoscopy implies overcoming new barriers. The light source is co-located with the endoscopes, hence moving and close to the body surfaces. The body surfaces have poor texture and abundant reflection due to fluids. The scene geometry includes a prevalent deformation. The video combines slow observation of areas of interest, with fast motions and long occlusions of the endoscope lenses.

It is our contribution the Endomapper dataset[3], which makes available, for the first time, **59 high quality calibrated recordings of complete routine endoscopies** (Fig. 1), making secondary use of medical data, i. e., just recording standard procedures that were going to be performed in any case, without any modification. Compared to ad-hoc recordings, secondary-use ones show realistic features and hence contain the actual challenges VSLAM will face in routine practice.

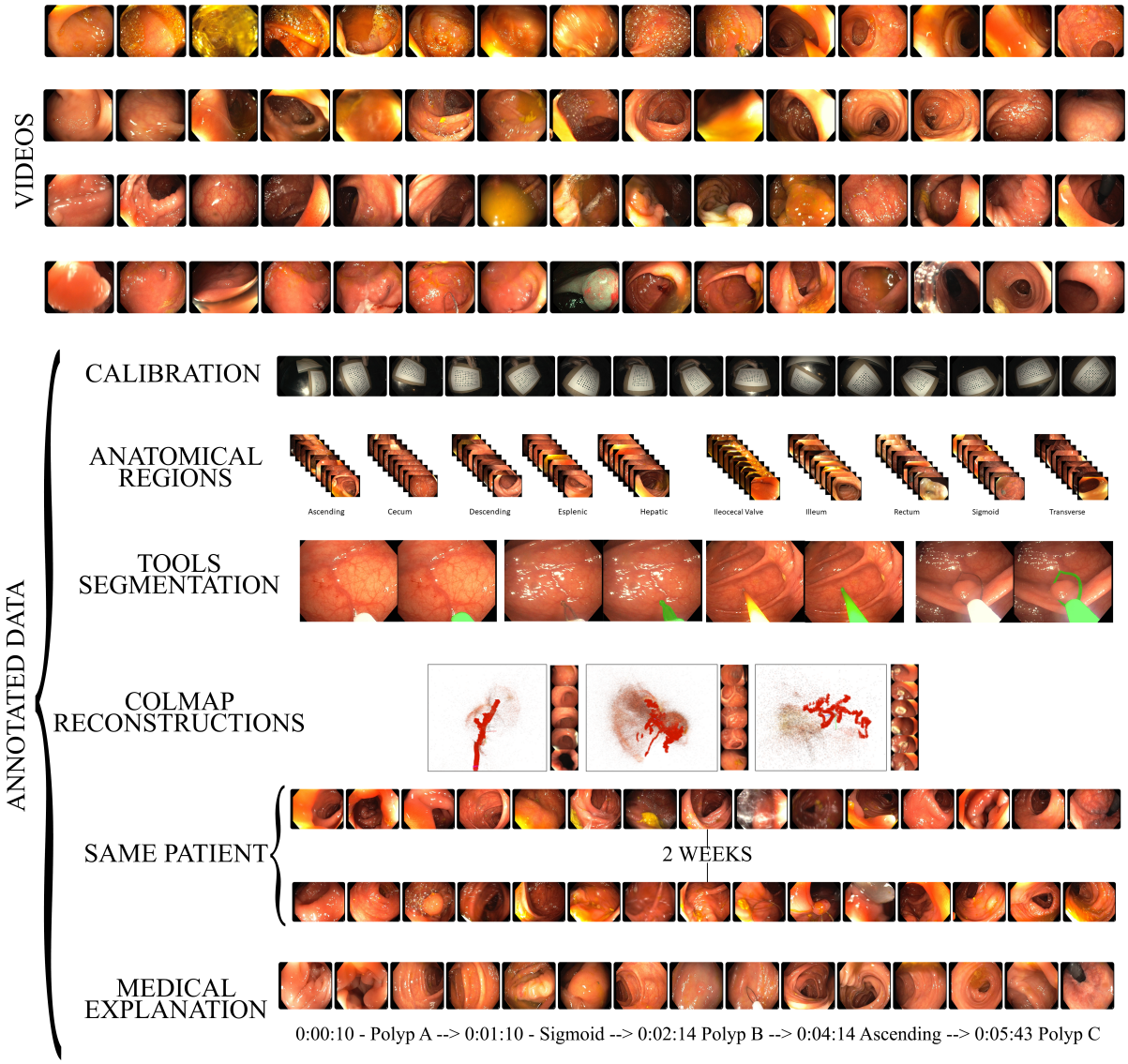


Figure 1. Overview of the Endomapper Dataset.

No other public dataset offers a comparable volume of full calibrated endoscopies in HD (see Table 1.) CVCClinicDB, GIANA and Kvasir focus on polyps detection, since they are often used as a CAD system. Other focus in segmentation, tools in Heilderberg or polyps in Kvasir-seg. However they only provide sparse image sets or short videos (less than 30 seconds). In contrast, we offer hours of real calibrated video, corresponding to the full procedures.

Endomapper includes colonoscopies, gastroscopies, and calibration videos along with photometric calibration parameters. There are screening colonoscopies with a thorough and slow exploration. To research map reuse in a second exploration, colonoscopies corresponding to the same patient but separated in time by several weeks are included.

Endomapper offers a sweet point of challenge, including easy video segments where traditional SfM or VLSAM algorithms work. In any case, all these methods fail at some point signaling what are the challenges to face. We can conclude that the dataset can spur research to identify and solve the challenges that VSLAM has to face in gastrointestinal environments.

Due to the monocular nature of the dataset, no ground truth geometry is available for quantitative evaluation. To address this issue, we include photorealistic sequences from a simulated colon, with ground truth geometry for the deforming scene and endoscopy trajectory.

Regarding metadata, some endoscopies include a description made by the endoscopist, in the form of text footage, of the procedure. The text describes the anatomical regions traversed, re-explorations of the same region, the performed interventions

Dataset	Purpose	Type of Data	Size of Dataset	Availability
CVC-ClinicDB[4]	Polyps segmentation	Images	612 images	Open Academic
Endoscopic artifact detection[5]	Artifact Detection	Images	5,138 images	Open Academic
GIANA 2021[6]	Polyp detection, segmentation and classification	Short Videos and Images	38 videos and 3000 images	By request
Kvasir[7]	Anatomical landmarks, Pathological findings, Therapeutic interventions and Quality of mucosal views	Images	4000 images	Open Academic
Kvasir-Seg[8]	Polpy segmentation	Images	8000 images	Open Academic
Nerthus[9]	Bowel preparation	Short Videos	21 videos (5525 frames)	Open Academic
Heilderberg[10]	Tools segmentation	Images	10040 images	By Request
HyperKvasir[11]	Anatomical landmarks, Pathological findings, Therapeutic interventions and Quality of mucosal views	Images & Short Videos	110079 images (10662 labeled) & 374 short videos	Open Academic
Endomapper (ours)	VSLAM	**Complete real endoscopies**	**20 videos (∼7.5 hours)**	By Request

Table 1. Overview of existing endoscopy datasets.

or the tools used. This footage indexes the videos to identify the interesting sections for SLAM.

Building on our dataset, the community can provide derived or metadata results to support subsequent research. Some examples of these derived data are included in the dataset: 1) Anatomical regions segmentation, at frame level, performed by a doctor after visualizing the video. 2) Tools segmentation in selected video sections, which can boost the tool segmentation performance in the endoscopy specific domain. 3) Structure from Motion (SfM) using COLMAP[12] which provides up to scale 6 DoF endoscope trajectory and 3D models for the video segments corresponding to smooth explorations of non-deforming scenes. The output of COLMAP can be used for the computer vision community as supervision for tasks such as image matching[13] or image retrieval[14].

2 Methods

The methodology used to create the dataset is explained in this section. First, a description of the recording procedure for the sequences in the dataset, including the description of the capture system and the type of recordings, is presented. Then, the description of the calibration procedure and the methodology used in both geometric and photometrical calibration is presented. Finally, there is a summary of the methods used to create each type of meta-data.

2.1 Recording endoscopies procedure

The acquisition of the sequences in the dataset was performed in the Hospital Clinico Universitario Lozano Blesa, Zaragoza (Spain), using an Olympus EVIS EXERA III CV-190 video processor, EVIS EXERA III CLV-190 light source, and EVIS EXERA III CF-H190 colonoscope or EVIS EXERA III GIF-H190 gastroscope.

The acquisition system is composed of a computer and a data acquisition card connected to the endoscopy tower via a Digital Visual Interface (DVI). Two different adquisition cards have been used: Epiphan Video DVI2USB 3.0 and Magewell Pro Capture DVI.

The video is grabbed at 1440×1080 at 40fps and 24RGBbits (Epiphan) or 1440×1080 at 50fps and 24RGBbits (Magewell). The videos are manually edited to remove any frame out of the body of the patient.

2.2 Calibration

The dataset uses 10 different colonoscopes and 8 different gastroscopes. The calibration sequences for all the colonoscopes and gastroscopes where acquired in a single session using a Lambertian pattern [1].

Figure 2 displays an example of the calibration videos, and the pattern. The Lambertian calibration pattern correspond to an array of circles pattern from the Vicalib[15] library. The physical size of the pattern used is $5,61 \times 9,82$ cm.

Geometric calibration

The calibration videos are processed by Vicalib[15] to obtain the endoscope intrinsic parameters according to Kannala & Brandt[16,17] model. The calibration defines eight intrinsic parameters: the four projective parameters (in pixels) f_x, f_y, C_x, C_y, and the four distortion coefficients k_1, k_2, k_3, k_4. We process 1 out of 20 frames and outlier matches are removed. Next, the

[1] Pattern obtained from calib.io]

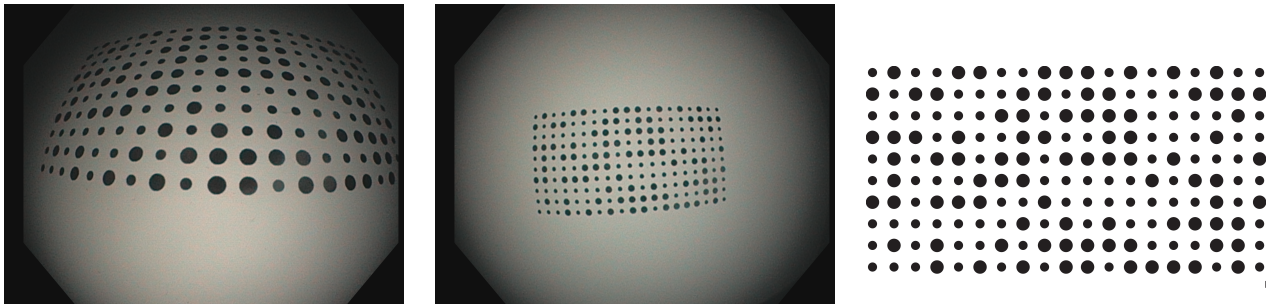


Figure 2. Examples of the Calibration images. Pattern used can be seeing in the third image.

projection model yielding the projection in pixels $\mathbf{u} = (u, v)$, for a 3D point with coordinates $\mathbf{X} = (x, y, z)$ with respect to the camera frame is described as:

$$u = f_x x_d + C_x, \qquad x_d = r_d \frac{x}{r} \tag{1}$$

$$v = f_y y_d + C_y, \qquad y_d = r_d \frac{y}{r} \tag{2}$$

where $r_d = \theta\left(1 + k_1\theta^2 + k_2\theta^4 + k_3\theta^6 + k_4\theta^8\right)$ is the distorted radius, $r = \sqrt{x^2 + y^2}$ is the undistored radius and $\theta = \arctan 2\,(r, z)$ is the angle between the incoming ray and the optical axis.

Photometric calibration

The light source and camera of the endoscope are calibrated to obtain a model able to reproduce the photometry of the recordings. Based on recent work about photometric reconstruction[18], we define a model that accounts for both the light emission and capture, as well as the interaction of the light with the geometry:

$$\mathscr{I}(\mathbf{u}) = \left(\frac{\mu(\mathbf{X})\ \sigma_o}{d^2}\ f_r(\omega_i, \omega_r)\ \cos\theta\ V(\mathbf{u})\ g_t \right)^{1/\gamma} \tag{3}$$

We denote as $\mathbf{X}$ a 3D point of the surface that appears in pixel $\mathbf{u}$ on the image. Thus the final pixel value $\mathscr{I}(\mathbf{u})$ depends on the light spread function $\mu(\mathbf{x})$ and the camera vignetting $V(\mathbf{u})$. The former models the light radiance going from the endoscope to the imaged world point. It acts as a multiplying factor of the maximum outgoing radiance σ_o. The latter accounts for the attenuation introduced by the camera lenses and capture system.

While light propagates, its radiance decreases as a function of the distance travelled d, following an inverse-square law. Then a bidirectional reflectance distribution function (BRDF) $f_r(\omega_i, \omega_r)$ defines how light is reflected from the surface to the camera. The projection of the light beam on the geometry introduces a cosine term of the angle θ between the incoming light ray ω_i and the surface normal. Finally, the endoscope applies an automatic gain g_t, that can vary at every t-th time instant, and a gamma curve ($\gamma = 2.2$) to improve the perceived dynamic range of the image.

In the endoscope, the distances between the light sources and the camera are small and mostly symmetrical. During calibration we assume that these sources can be modelled as a single light located at the camera optical centre. Both the light spread function and the camera vignetting are jointly estimated assuming radial symmetry, following an off-axis cosine fall-off:

$$\mu'(\mathbf{X}) = \mu(\mathbf{X}) \cdot V(\mathbf{u}) = \cos^k \alpha \tag{4}$$

where the off-axis angle α is the angle between the projection ray and the camera forward, and k is a parameter of the model.

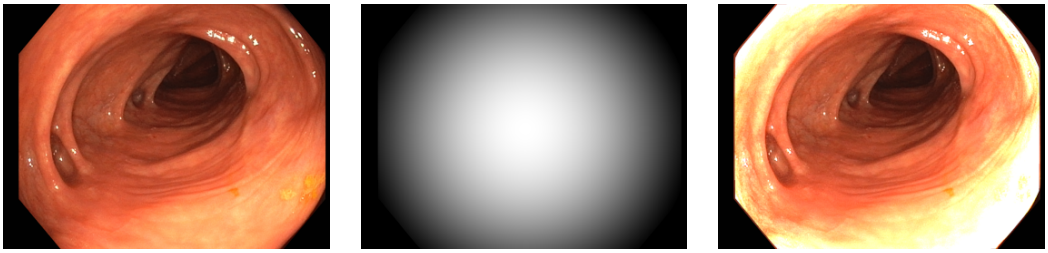
The parameters of the model are estimated by optimising a photometric loss on the white areas of the Vicalib pattern. The results of the calibration (Fig. 3) provide a 2D weighting of the photometric effects caused by the vignetting and the light spread function. This can be used to compensate those effects and recover the real radiance on the scene.

2.3 Simulated colon

VR-Caps[19] simulator is used to generate photerealistic synthetic image sequences of a 3D colon model obtained from a Computed Tomography. Since this is a simulation, we have access full to scene configuration:camera calibration, deformations, trajectory and illumination, hence to the and ground truth geometry,camera pose and 3D deforming scene.

For the same endoscope trajectory, we generate different sequences with more aggressive deformations to allow ablative studies with respect the deformation magnitude. Deformations applied are described by the next equation:

$$V_y^t = V_y^0 + A\sin(\omega t + V_x^0 + V_y^0 + V_z^0) \tag{5}$$



Original image | Vignetting & $\mu(\mathbf{x})$ | Compensated image

Figure 3. Example of photometric calibration results.

where V_x^0, V_y^0 and V_z^0 are the coordinates of the surface point at rest. We can control the magnitude and velocity of the deformations according to the parameters A and ω respectively, which corresponds to the maximum excursion and velocity of the deformations respectively. We also modify the colon texture to increase its contrast.

2.4 Meta-data

For a set of selected recordings several types of meta-data useful for Visual SLAM are provided. In this subsection, a description of the meta-data and the methodology for its acquisition is presented.

Text footage

The description of procedure made by the endoscopist is registered during the exploration. It includes the anatomical regions traversed, the interventions, the medical findings such as polyp approximated size, the tools used or the sections with NBI (Narrow-band imaging) illumination. This description is made available as synchronized text footage with the videos. This metadata can be useful to identify the sections of the video more promising for SLAM, the reobservations of the same region or the interaction with tools of known size.

Anatomical regions

Anatomical section recognition is useful to create topological maps of the colon. These maps can be used to create smaller reconstructions with less probability of error. Some colonoscopies procedures were annotated by a medical staff of the project after the recording. Multiple careful visualizations were necessary to delimit the ten anatomical regions summarized in Fig. 4.

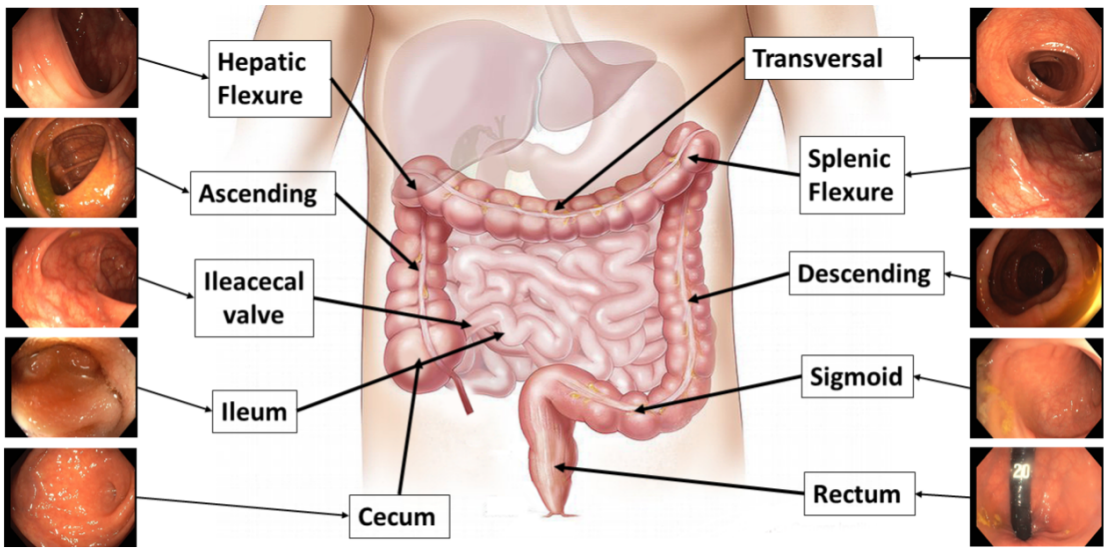


Figure 4. Explanation of the anatomical places labeled.

Tools segmentation

Tool segmentation is one of the challenges in colonoscopies for AI. Since they occlude and mislead other algorithms, many works try to mask them. Tools were manually segmented using Odin CAT tool[20], which allows to maintain a mask between

frames, giving a more robust annotation.

COLMAP 3D reconstruction

Traditional SIFT based rigid SfM algorithms are able to process and produce partial reconstructions from the colonoscopies. We include some examples of the output of COLMAP[12,21] processing in our sequences, which provides a first approximation for the up to scale camera trajectory and the scene's sparse structure. This information can be organized to produce weak supervision in the form of sparse depth maps, local correspondences between frames, image-to-image labels (frames depicting the same place) or relative camera pose transformation between frames. Several computer vision tasks like depth prediction, image matching, image retrieval and visual localization can greatly benefit from this kind of supervision. Megadepth[22] is a well-known dataset that uses this SfM procedure to obtain 3D point clouds, similar to us. It is being extensively used for deep learning supervision[13,23,24]. Other works employed SfM to identify co-visible frames in the recordings, which has proven to be useful to train CNNs for place recognition in landmark images[14] and in colonoscopy sequences[25,26].

For our recordings, we apply exhaustive guided matching between all the images in the sequence to associate frames far away in time. We use our camera calibration and we do not optimize it during the COLMAP bundle adjustment. The minimum triangulation angle is relaxed to 8 degrees during the initialization of the models. The rest of parameters are left as default.

Recordings from the same patient

One of the main obstacle in colon reconstruction is consistency between colonoscopies in longitudinal studies. Thanks to the colorectal cancer screening program, colonoscopy pairs from the same patient were registered. This would help to evaluate lifelong capabilities of the developed visual SLAM algorithms.

3 Data Records

This section describes the dataset structure and details the meta-data available. A summary of the dataset structure can be seen in Fig. 5. At publication time, there is a total of 59 sequences and the length of the sequences goes from less than ten minutes to more than half an hour. File `DatasetSummary.xls` contains the full list of sequences in the dataset up-to-date.

3.1 Video recordings

Each procedure has a directory `Sequence_XXXX` (`XXXX` is the sequence number) that contains:

1. The directory `meta-data` that contains all the meta-data files associated to the sequence. These files are described in the next section.
2. The recording, `Sequence_XXXX.mov` where the actual recording is. The video codec is H264[27], a lossy compression but with a profile of High 4:4:4 with 4.2 level and a bit rate of 7Mbps. It offres an optimal size vs. quality trade-off for lossy compression.
3. The thumbnail version, `Sequence_XXXX_thumbnail.webm` contains a compressed version of the recording for easy visualization. This version uses the free codec `libvpx`[28], at 320×240 resolution.
4. A subtitle file, `Sequence_XXXX.srt` if the video has text footage in the form of text subtitles.
5. The metadata file, `Seq_XXXX_info.json`, where sequence number, endoscope number and the type of metadata of the procedure is stored.

A folder `Lossless_sequences` contains also the lossless version of the videos. This version uses codec ffv1 version 3 with a bitrate of 310 Mbps.

3.2 Camera calibration

Calibration videos are identified by the endoscope number. There is a directory `Endoscope_XXXX` (`XXXX` is the endoscope number), for each endoscope that contains:

1. The calibration video `Endoscope_XX.mov` including the actual calibrations recording. This version is the lossy H264 version and the lossless version can be found the lossless folder metioned before.
2. The Geometric Calibration `Endoscope_XX_geometrical.xml` containing the results of the geometrical calibration.
3. The Photometrical Calibration `Endoscope_XX_photometrical.xml` containing the results of the photometrical calibration.

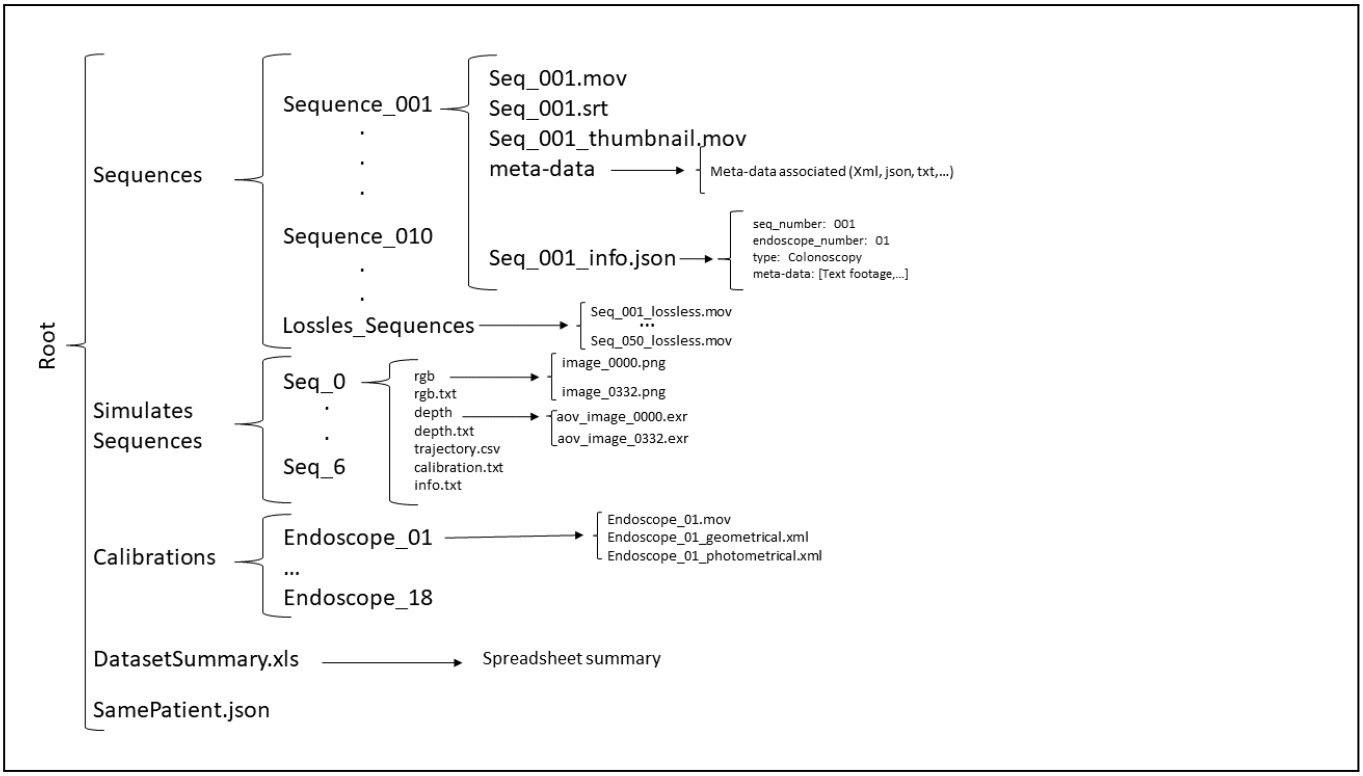


Figure 5. Directory structure of the dataset. The right part is an example of the information in the `info.json` for `Sequence_001`

Geometric Calibration

The file `Endoscope_XX_geometrical.xml` is the output calibration from the Vicalib[15]. This XML file contains the intrinsic parameters of the camera $(f_x, f_y, C_x, C_y, k_1, k_2, k_3, k_4)$ following the Vicalib output format.

Photometric Calibration

The photometric calibration file, `Endoscope_XX_photometrical.xml`, contains the calibrated parameters of the light source and the camera of the endoscope. An endoscope's `<rig>` may have one or more `<camera>` tags, associated with one or more `<light>` sources. Currently, only a single camera and a single virtual light are supported.

Each camera tag has a particular `<camera_model>`. The model type "`photodepth_gamma`" has a single parameter, the value of the gamma γ response function in Eq. (3). Regarding the light source, the `<light_model>` type "`photodepth_L_spread`" has two parameters, one for the maximum outgoing radiance σ_o, and another one k, for the cosine fall-off of the light. In addition, the `<T_cl>` tag specifies the light source's pose with respect to the first camera.

3.3 Simulated colon

There is a directory, `seq_X` (`X` is the sequence number) per each sequence obtained from simulation. The directory contains:

1. The directory `rgb` containing the RGB images of the sequence in `png` format.
2. The directory `depth` containing depth images for each RGB image of the sequence stored in `exr` format.
3. A file `rgb.txt` with a list of file names of all RGB images of the sequence.
4. A file `depth.txt` with a list of the file names of all depth images of the sequence.
5. A file `trajectory.csv` containing the ground truth camera trajectory.
6. A file `calibration.txt` containing the simulated camera calibration.
7. A file `info.txt` containing the deformations applied, its parameters and units.

3.4 Meta-data

This section describes the records and formats for each type of meta-data.

Text footage

Two files: `text_footaje_XXX.json` and `text_footage_XXX.srt` are included inside the folder `meta-data`. The `.json` file contains an structure with the timestamp and the associated text. The text footage is also included in `.srt` format to ease the visualization synchronized with the video. The references to identify the tools used during the procedure are stored in the `meta-data` directory.

Anatomical regions

Table 2 shows the detailed number of frames labelled for each region in each video. The dataset contains this information in a file named `Anatomical_Regions_XXXX.txt` with the format `Frame###;region label;` in each line.

Tool Segmentation

There are 3422 frames with tools segmented across two different colonoscopies as detailed in Table. 3. The segmentations for each video can be found in file `tool_segmentation_XXX.xml`. This file contains, for each segmented frame, the id of the frame and list of 2D points coordinate that define the tool segmentation as a binary polygon. Some examples can be seen in Fig. 6.

COLMAP 3D reconstruction

The reconstructions obtained using the parameters explained before are shown in Table. 4. The reconstructions are stored in the dataset following the text format of COLMAP[2]. Fig. 7 shows some examples of these reconstructions.

Same patient recordings

A file `SamePatient.json` is stored in the root folder containing which sequences are from the same patient and the time that separate both sequences.

Sections	Total Frames	rectum	sigmoid	descending	esplenic	transverse	hepatic	ascending	ileocecal	ileum	cecum
Seq_003	39220	760	1420	3820	160	1720	920	29360	160	900	0
Seq_011	12920	740	5960	2220	1100	1660	1240	0	0	0	0
Seq_012	16680	180	1360	1800	1100	2100	1180	6700	760	0	1500
Total	68820	1680	8740	7840	2360	5480	3340	36060	920	900	1500

Table 2. Summary of the anatomical sections per video and label.

Sequence	Nº Frames segmented
Seq_003	3168
Seq_012	254

Table 3. Summary of the frames with tool segmentation.

Sequence	Total frames	Frames reconstructed	Clusters obtained
Seq_001	14824	5809(39,18%)	50
Seq_002	23375	8133(34,79%)	50

Table 4. Summary of Colmap 3D reconstruction.Second column is the number of frames in the video and third column is the number of frames used for the reconstruction.

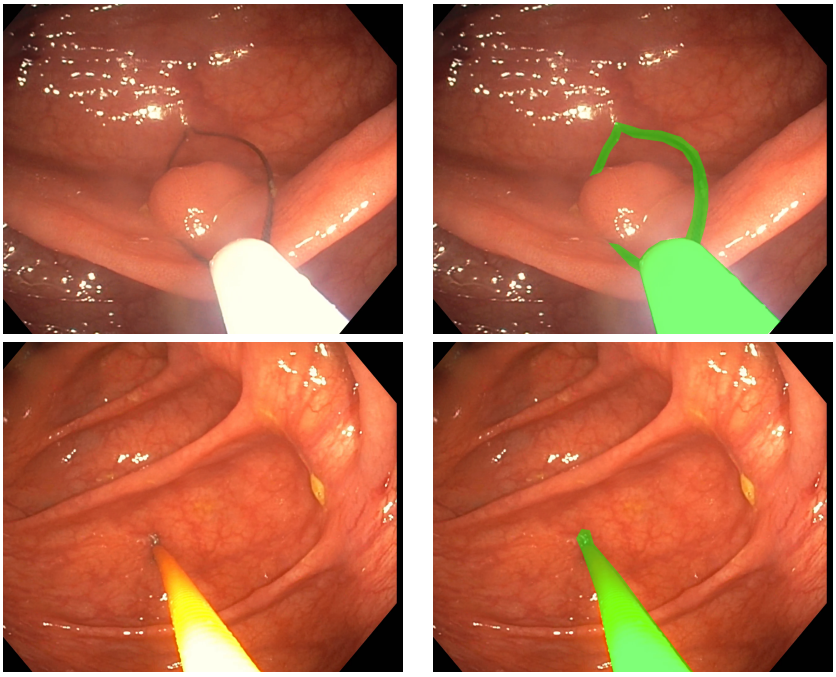
4 Usage Notes

The dataset is available on the Synapse platform[3]. The access will be restricted to registered users affiliated to a non-profit organization.

References

1. Campos, C., Elvira, R., Rodríguez, J. J. G., Montiel, J. M. & Tardós, J. D. ORB-SLAM3: An accurate open-source library for visual, visual–inertial, and multimap SLAM. *IEEE Transactions on Robotics* **37**, 1874–1890 (2021).
2. Engel, J., Koltun, V. & Cremers, D. Direct sparse odometry. *IEEE Transactions on Pattern Analysis Mach. Intell.* (2017).

[2] https://colmap.github.io/format.html



Original image Tool masked

Figure 6. Examples from the tool segmentation mask. These examples are from sequence 9.

3. Azagra, P. *et al.* Endomapper dataset a complete endoscopy procedures dataset. https://doi.org/10.7303/syn26707219, 10.7303/SYN26707219 (2022). Synapse.
4. Fernández-Esparrach, G. *et al.* Exploring the clinical potential of an automatic colonic polyp detection method based on the creation of energy maps. *Endoscopy* **48**, 10.1055/s-0042-108434 (2016).
5. Ali, S. *et al.* An objective comparison of detection and segmentation algorithms for artefacts in clinical endoscopy. *Sci. Reports* **10**, 10.1038/s41598-020-59413-5 (2020).
6. Bernal, J., Tudela, Y., Riera, M. & Sánchez, F. J. *Polyp Detection in Colonoscopy Videos*, 163–169 (Springer International Publishing, Cham, 2021).
7. Pogorelov, K. *et al.* Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. 10.1145/3083187.3083212 (2017).
8. Jha, D. *et al.* Kvasir-seg: A segmented polyp dataset. In *International Conference on Multimedia Modeling*, 451–462 (Springer, 2020).
9. Pogorelov, K. *et al.* Nerthus: A bowel preparation quality video dataset. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, MMSys'17, 170–174, 10.1145/3083187.3083216 (Association for Computing Machinery, New York, NY, USA, 2017).
10. Maier-Hein, L. *et al.* Heidelberg colorectal data set for surgical data science in the sensor operating room. *Sci. data* **8**, 1–11 (2021).
11. Borgli, H. *et al.* Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Sci. Data* **7**, 1–14 (2020).
12. Schönberger, J. L. & Frahm, J.-M. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2016).
13. Dusmanu, M. *et al.* D2-net: A trainable cnn for joint description and detection of local features. In *Proceedings of the IEEE/cvf Conference on Computer Vision and Pattern Recognition*, 8092–8101 (2019).
14. Radenović, F., Tolias, G. & Chum, O. Fine-tuning cnn image retrieval with no human annotation. *IEEE Transactions on Pattern Analysis Mach. intelligence* **41**, 1655–1668 (2018).
15. (n.d.). Vicalib library. https://github.com/arpg/vicalib ((n.d.)). (accessed: 14.10.2020).

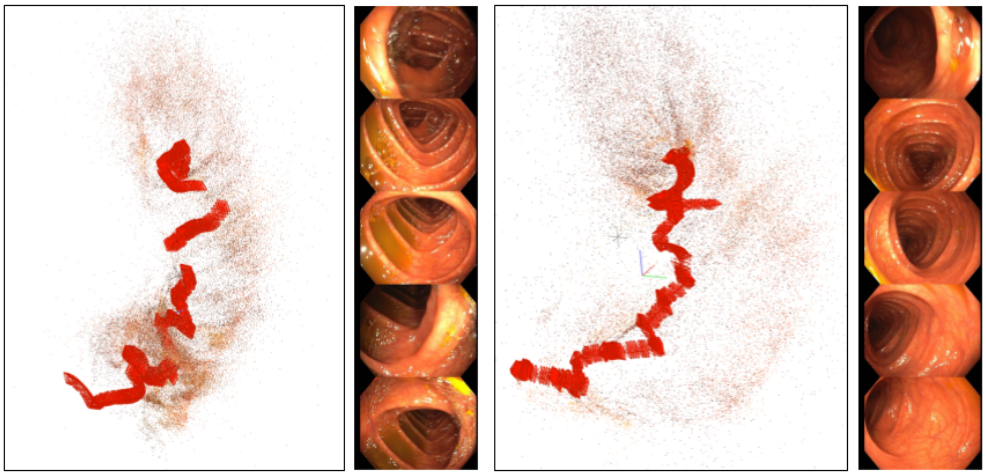


Figure 7. Examples from the COLMAP reconstructions obtained.

16. Kannala, J. & Brandt, S. S. A generic camera model and calibration method for conventional, wide-angle, and fish-eye lenses. *IEEE Transactions on Pattern Analysis Mach. Intell.* **28**, 1335–1340 (2006).

17. Usenko, V., Demmel, N. & Cremers, D. The double sphere camera model. In *2018 International Conference on 3D Vision (3DV)*, 552–560 (IEEE, 2018).

18. Hao, Y. *et al.* Photometric stereo-based depth map reconstruction for monocular capsule endoscopy. *Sensors* **20**, 5403 (2020).

19. İncetan, K. *et al.* Vr-caps: a virtual environment for capsule endoscopy. *Med. Image Analysis* **70**, 101990 (2021).

20. Odin. Odin cat tool. https://cat-aws.odin-vision.com.

21. Schönberger, J. L., Zheng, E., Pollefeys, M. & Frahm, J.-M. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)* (2016).

22. Li, Z. & Snavely, N. Megadepth: Learning single-view depth prediction from internet photos. In *Computer Vision and Pattern Recognition (CVPR)* (2018).

23. Sarlin, P.-E. *et al.* Back to the feature: Learning robust camera localization from pixels to pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3247–3257 (2021).

24. Yang, T.-Y., Nguyen, D.-K., Heijnen, H. & Balntas, V. Ur2kid: Unifying retrieval, keypoint detection, and keypoint description without local correspondence supervision. *arXiv preprint arXiv:2001.07252* (2020).

25. Morlana, J., Millán, P. A., Civera, J. & Montiel, J. M. Self-supervised visual place recognition for colonoscopy sequences. In *MIDL* (2021).

26. Ma, R. *et al.* Colon10k: A benchmark for place recognition in colonoscopy. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, 1279–1283, 10.1109/ISBI48211.2021.9433780 (2021).

27. Richardson, I. E. *H. 264 and MPEG-4 video compression: video coding for next-generation multimedia* (John Wiley & Sons, 2004).

28. Grange, A. & De Rivaz, P. VP9 bitstream & decoding process specification. *Google* .

Acknowledgements

This work was supported by EU-H2020 grant 863146: ENDOMAPPER.

Author contributions statement

J.M.M.M. originated the concept of dataset. P.A., L.R., J.C., J.D.T., A.C.M. and J.M.M.M. designed the dataset details. C.S., A.F., and A.L. performed the endoscopies, provided medical explanations and anatomical labels. P.A., L.R. and C.O. J.M.M.M. designed and operated the data acquisition system and created the database. P.A., V.M.B., J.D.T. and J.M.M.M. performed

endoscope's calibration. J.J.G.R. and J.D.T. provided colon simulations. O.L.B., J.M. and J.M.M.M. provided COLMAP reconstructions. C.T., L.R. and A.C.M provided tool segmentation. P.A. and J.L. provided anatomical landmark annotations. P.A., L.R., O.L.B., C.T., J.M., D.R., V.M.B., J.J.G.R., R.E., J.C., J.D.T., A.C.M. and J.M.M.M performed the analysis and technical validation. P.A., L.R., J.M., V.M.B., J.J.G.R., J.D.T. and J.M.M.M. created and edited the manuscript. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.