Tracking Adaptation to Improve SuperPoint for 3D Reconstruction in Endoscopy

O. León Barbed[1(✉)], José M. M. Montiel[1], Pascal Fua[2], and Ana C. Murillo[1]

[1] DIIS-i3A, University of Zaragoza, Zaragoza, Spain
leon@unizar.es
[2] CVLAB, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

Abstract. Endoscopy is the gold standard procedure for early detection and treatment of numerous diseases. Obtaining 3D reconstructions from real endoscopic videos would facilitate the development of assistive tools for practitioners, but it is a challenging problem for current Structure From Motion (SfM) methods. Feature extraction and matching are key steps in SfM approaches, and these are particularly difficult in the endoscopy domain due to deformations, poor texture, and numerous artifacts in the images. This work presents a novel learned model for feature extraction in endoscopy, called SuperPoint-E, which improves upon existing work using recordings from real medical practice. SuperPoint-E is based on the SuperPoint architecture but it is trained with a novel supervision strategy. The supervisory signal used in our work comes from features extracted with existing detectors (SIFT and SuperPoint) that can be successfully tracked and triangulated in short endoscopy clips (building a 3D model using COLMAP). In our experiments, SuperPoint-E obtains more and better features than any of the baseline detectors used as supervision. We validate the effectiveness of our model for 3D reconstruction in real endoscopy data. Code and model: https://github.com/LeonBP/SuperPointTrackingAdaptation.

Keywords: deep learning · structure from motion · local features · endoscopy

1 Introduction

Endoscopy is an important medical procedure with many applications, from routine screening to detection of early signs of cancer and minimally invasive treatment. Automatic analysis and understanding of these videos raises many opportunities for novel assistive and automatization tasks on endoscopy procedures. Obtaining 3D models from the intracorporeal scenes captured in endoscopies is an essential step to enable these novel tasks and build applications,

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43907-0_56.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14220, pp. 583–593, 2023.
https://doi.org/10.1007/978-3-031-43907-0_56

for example, for improved monitoring of existing patients or augmented reality during training or real explorations.

3D reconstruction strategies have been studied for long, and one crucial step in these strategies is feature detection and matching which serves as input for Structure from Motion (SfM) pipelines. Endoscopic images are a challenging case for feature detection and matching, due to several well known challenges for these tasks, such as lack of texture, or the presence of frequent artifacts, like specular reflections. These problems are accentuated when all the elements in the scene are deformable, as it is the case in most endoscopy scenarios, and in particular in the real use case studied in our work, the lower gastrointestinal tract explored with colonoscopies. Existing 3D reconstruction pipelines are able to build small 3D models out of short clips from real and complete recordings [1]. One of the current bottle-necks to obtain better 3D models is the lack of more abundant and higher quality correspondences in real data.

This work introduces SuperPoint-E, a new model to extract interest points from endoscopic images. We build on the well known SuperPoint architecture [5], a seminal work that delivers state-of-the-art results when coupled with downstream tasks[1]. Our main contribution is a novel supervision strategy to train the model. We propose to automatically generate reliable training data from video sequences by tracking feature points from existing detection methods, which do not require training. We select *good features* with the COLMAP SfM pipeline [21], generating training examples with feature points that can be tracked across several images according to COLMAP result. When used to train SuperPoint, our approach yields a self-supervised method outperforming current ones.

2 Related Work

3D reconstruction is an open problem for laparoscopic and endoscopic settings [14] of high interest for the community. This idea is supported for example by recent efforts on collecting new public dataset, to further advance in this field, such as endoscopic recordings from EndoSLAM [16] and EndoMapper [1] datasets. Earlier works like Grasa *et al.* [7] have evaluated the performance of modern SLAM approaches on endoscopic sequences. Mahmoud *et al.* [13] improved the performance of such methods in laparoscopic sequences. More recent approaches attempt to tackle specific endoscopy challenges, such as the deformation [18] or the artifacts due to specular reflections in the feature extraction step [2].

Well known SfM and SLAM pipelines rely on accurate and robust feature extraction methods. COLMAP [21,22], a public SfM tool, uses SIFT [11] features while ORB-SLAM [15] extracts ORB [19] features because of their efficiency. Both these feature extraction methods count with classical, hand-crafted descriptors that allowed to build such complex applications. However, transferring that performance to endoscopy settings remains a difficult task due to

[1] https://www.cs.ubc.ca/research/image-matching-challenge/2021/leaderboard/.

several challenges. Artifacts or the lack of texture result in low amount of correspondences along real endoscopy videos, what motivates the need for improved strategies.

Deep learning methods for feature extraction and matching is a very active research field. The survey Ma *et al.* [12] shows the introduction of deep learning methods to feature detection and matching. Notable mentions are SuperPoint [5] for its self-supervised approach, R2D2 [17] for using reliability metrics as output of the network instead of the features themselves and D2-Net [6] that built a describe-and-detect strategy that aims to improve SfM applicability. Exporting this progress to the matching stage, DISK [24] proposes a formulation of the problem to optimize in an end-to-end manner. Other recent works have extended the networks to take advantage of the advances in attention for the matching task, as in SuperGlue [20] and LoFTR [23].

In this work we improve the performance of SuperPoint [5] on endoscopy images. We chose SuperPoint because it is a seminal work that has inspired many follow up works, and it is still among the top performers on current feature matching challenges [10]. Similar to DeTone *et al.* [4], we explore improvements on feature extraction that provide good properties for downstream tasks. They design an end-to-end method to optimize the visual odometry computed with their features. Differently, we propose to supervise our training with points that have been successfully used for 3D reconstruction using existing SfM pipelines. With this supervision, we train a model able to extract more features with good properties for SfM algorithms, e.g., being spread and out of large specularities.

3 *Tracking Adaptation* for Local Feature Learning

Superpoint supervision is referred to as *Homographic Adaptation* and assumes that the surfaces are locally plane, which is not the case in our data. Instead, we propose to use 3D reconstructions of points tracked along image sequences. This makes no assumptions about the local surface shapes and we will show in Sect. 4 that this yields a better trained network. We will refer to this as *Tracking Adaptation* and we will here describe how we obtain the tracks.

SfM as Supervision for Feature Extraction. We generate examples of *good features* by identifying features that were successfully reconstructed with existing methods for each sequence in our training set. Our training set contains short sequences (4–7 s) from the complete colonoscopy recordings in EndoMapper dataset where COLMAP software was able to obtain a 3D reconstruction. This is a very challenging domain, and existing SfM pipelines fail in longer videos.

3D Reconstruction of Training Set Videos. We generate 3D reconstructions for all our training sequences with out-of-the-box COLMAP. In particular, we use the following blocks: *feature_extractor*, *exhaustive_matcher* and *mapper*. Configuration parameters are detailed in the supplementary materials. We turn on the "guided_matching" option for the *exhaustive_matcher* module to find the best matches possible. We additionally compute the 3D reconstruction for the

same sequences with a modified COLMAP pipeline that uses the official SuperPoint and SuperGlue[2] implementation with the *indoor* set of weights. All the parameters are left as default except for the keypoint_threshold= 0.015 and the nms_radius= 1. After providing the SuperGlue resulting matches to COLMAP, we execute only the *mapper* module with the same configuration as before.

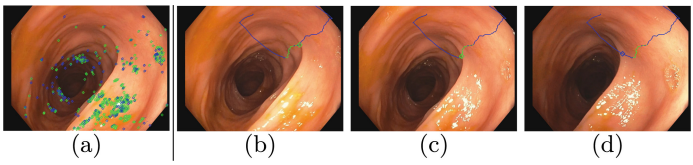

Fig. 1. Supervision points obtained from a COLMAP reconstruction. (a) All 3D points are reprojected into each video frame. We distinguish points that were originally detected in this frame (green) and points that were not (blue). (b–d) analyze a complete point track, i.e., all the positions of the same 3D point along the sequence. The *reliable track* for this point is the green segment. (b) The track starts when a point is first detected. (c) Movement of the point along the video. (d) When the feature is not detected anymore (e.g., because of occlusion), it is depicted in blue from then on. (Color figure online)

Re-project Good Features to the Training Set Frames. A successful 3D reconstruction includes the computed positions of the cameras that took the images and a point cloud with 3D coordinates of the triangulated points. We use the camera poses, the points' coordinates and the camera calibration parameters to reproject the 3D point cloud points into every image. Not all points were originally detected and triangulated at all frames, so we establish two types of reprojected points. If they were "originally" detected and matched in a particular image, we set them to green. Otherwise, we set them to blue (see Fig. 1).

For supervision, we only use reprojected points that fall within a *reliable track*. A *reliable track* is an interval bounded by green points. So, the reprojected points selected for training are either green or have preceding and subsequent green points along its track.

The different appearances of the same 3D point in different frames of the *track* are our **correspondences** for training our models. Figure 1 contains examples of reprojected points and an example of a *reliable track*.

Deep Feature Extraction for Endoscopy. SuperPoint uses a fully-convolutional network as backbone and learns to extract good features using *homographic adaptation*: extracting features that are robust to homographic deformations. It achieves this by using as supervision Y the average detections over several random homographic deformations of the same image. The feature extraction network then is run on an image I and a warped version I' of it with a new

[2] https://github.com/magicleap/SuperGluePretrainedNetwork.

homography. The network optimizes the loss function

$$\mathcal{L}_{SP}\left(\mathcal{X},\mathcal{X}',\mathcal{D},\mathcal{D}';Y,Y',S\right)=\mathcal{L}_p\left(\mathcal{X},Y\right)+\mathcal{L}_p\left(\mathcal{X}',Y'\right)+\lambda\mathcal{L}_d\left(\mathcal{D},\mathcal{D}',S\right), \quad (1)$$

where $\mathcal{X}$ and $\mathcal{D}$ are the detection and description heads' outputs, respectively. Y is the supervision for the detection. S is the correspondence between I and I' computed from the homography. $\mathcal{L}_p$ is the detection loss that measures the discrepancies between the supervision Y and the detection head's output $\mathcal{X}$. $\lambda = 1$ is a weighting parameter. $\mathcal{L}_d$ is the description loss that measures the discrepancies between both description head's outputs $\mathcal{D}$ and $\mathcal{D}'$ using S.

Using our new supervision from SfM in the form of *tracks* of points, we propose a new loss to train SuperPoint that is more aligned with our goal, called *tracking adaptation*. Instead of an image I and a warped version I', we use different images I_a and I_b from the same sequence. The supervision Y for the detection in this case is the set of points that have been reprojected on I_a and I_b from the 3D reconstruction. The detection loss $\mathcal{L}_p$ is calculated as in the original SuperPoint. We replace the description loss $\mathcal{L}_d$ for a new tracking loss

$$\mathcal{L}_t\left(\mathcal{D}_a,\mathcal{D}_b,\mathcal{T}\right)=\frac{1}{|\mathcal{T}|^2}\sum_{i=1}^{|\mathcal{T}|}\sum_{j=1}^{|\mathcal{T}|}l_t\left(\mathbf{d}_{a_i},\mathbf{d}_{b_j},i,j\right), \quad (2)$$

$$\text{with} \qquad l_t\left(\mathbf{d}_{a_i},\mathbf{d}_{b_j},i,j\right)=\left\{\begin{matrix}\lambda_t\max(0,m_p-\mathbf{d}_{a_i}^T\mathbf{d}_{b_j}) & \text{if } i=j,\\ \max(0,\mathbf{d}_{a_i}^T\mathbf{d}_{b_j}-m_n) & \text{if } i\neq j\end{matrix}\right\}, \quad (3)$$

where $\mathcal{D}_a$ and $\mathcal{D}_b$ are the description head's outputs for I_a and I_b, respectively. $\mathcal{T}$ is the set of all the *tracks* that appear in both images. l_t is a common triplet loss that measures the distance between positive pairs (weighting parameter $\lambda_t = 1$ and positive margin $m_p = 1$) and the distance between negative pairs (negative margin $m_n = 0.2$). Two descriptors from different images $\mathbf{d}_{a_i}$ and $\mathbf{d}_{b_j}$ are a positive pair if they belong to the same *track* ($i = j$), and negative pair otherwise ($i \neq j$).

4 Experiments

The following experiments demonstrate the proposed feature detection efficacy to obtain 3D models on real colonoscopy videos, comparing different variations of our approach and relevant baseline methods.

Dataset. We seek techniques that are applicable to real medical data, so we train and evaluate with subsequences from the EndoMapper dataset [1], which contains a hundred complete endoscopy recordings obtained during regular medical practice. We use COLMAP 3D reconstructions obtained from subsequences from this dataset (11260 frames from 65 reconstructions obtained along 14 different videos for training, and 838 frames from 7 reconstructions from 6 different videos for testing). The exact details are in the supplementary material.

Baselines and our Variations. We use COLMAP as our first baseline. It uses **SIFT** features and a standard guided matching algorithm to produce very accurate camera pose estimates. We also include as baseline the results of SuperPoint (**SP**) with SuperGlue matches and the COLMAP reconstruction module. The configuration for both baselines is the same as detailed in Sect. 3. We evaluate different variations of the original SuperPoint. All models were trained with a modification of a PyTorch implementation of SuperPoint [9]. Training parameters in supplementary material. The models differ in the supervision used and the loss applied in the training, as detailed in the first four columns of Table 1.

Table 1. Ablation study. Configuration of the training (left), and average reconstruction results, i.e., quality metrics (right). Best results highlighted in bold.

	Supervision & Train Config.			Reconstruction Test Results				
	point	*match*	*loss*	$\|\|3DIm\|\|$	$\|\|3DPts\|\|$	Err	Err-10K	len (Tr)
SP [5]	SP-O	H	SP (original)	93.9%	6421.3	**1.47**	1.47	6.86
SP-E v0	SF*	H	SP (original)	97.3%	12707.9	1.66	1.50	8.39
SP-E v1	SF*	TR	tr-2	98.6%	13255.1	1.69	1.51	8.95
SP-E v1	SF*+SP*	TR	tr-2	99.1%	28308.3	1.74	1.13	9.45
SP-E v2	SF*+SP*	TR	tr-N	99.1%	**34838.0**	1.75	**1.02**	9.53
SP-E v3	SF*+SP*	H + TR	SP + tr-N	**99.2**%	30777.6	1.74	1.09	**9.65**

point (Base point detector): SP-O: original superpoint detector; SF*/SP*: SIFT/SP points that were successfully reconstructed after the COLMAP optimization, reprojected in each video frame.
match (Matches Supervision): H: Homography based, i.e., *Homographic adaptation* from original SuperPoint work; TR: The proposed *Tracking adaptation*.
loss (Loss used for training): SP: original SuperPoint training loss; Tr-2 or Tr-N: *track*-based loss. Tr-2 means that the loss is computed for every pair of images in the track. Tr-N means we optimize simultaneously N views of the track (N=4 in our experiments).

Ablation Study. Table 1 (last five columns) summarizes the performance of our approach variations. We run all the models on the **Test set** subsequences to extract points. Matches between the points in two images are obtained with bi-directional nearest neighbor algorithm with L2 distance. Points and matches are given to COLMAP and the *mapper* module (configuration in supplementary material) attempts to generate a 3D reconstruction. The reconstruction quality statistics used to illustrate the performance of each detector are:

- $\|3DIm\|$: **Fraction of images** from the subsequence successfully introduced in the reconstruction. The closer to 100% the better.
- $\|3DPts\|$: **Number of points** that were successfully reconstructed. The more points the better, since it means a denser coverage of the scene.
- Err: Mean **reprojection error** of the 3D points after being reprojected onto the images of the subsequence.
- Err-10K: Mean **reprojection error of the best 10000** points of the reconstruction. Since all reconstructions have outliers that skew the average, this metric is more representative of the performance of the models.

– len(Tr): Mean **track length** represents the average number of images where a point is being consecutively matched, tracked.

SP-E v2 (SP-E moving forward) is our best variation, with the highest amount of reconstructed points and the lowest reprojection error for top 10000.

SfM Results Comparison. This experiment compares the performance of the considered baselines against the best configuration of our feature extraction model. Table 2 contains a summary of the results. In most metrics we observe a significant improvement using SP-E compared to the others. For example, the number of points at the final reconstruction is more than three times higher (see

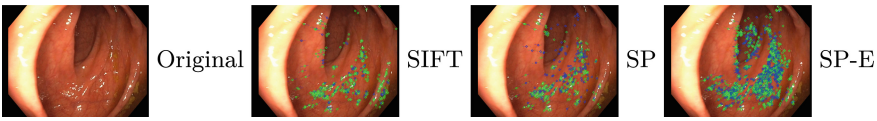

Fig. 2. Example of the points reconstructed by each method. Each point in each image has been reconstructed after the corresponding COLMAP reconstruction process.

Table 2. Reconstruction quality metrics for the comparison to the baselines.

Subsequence	001_1	002_1	014_1	016_1	017_1	095_1	095_2	Avg	(Std)
Reconstructed images ($\|3DIm\|$)									
Total$^+$	107	155	109	119	125	118	105	119.7	(15.9)
SIFT	98.1%	91.6%	71.6%	**100%**	52.0%	97.5%	99.0%	87.1%	(17.0)
SP	**100%**	**100%**	**93.6%**	**100%**	89.6%	99.2%	**100%**	97.5%	(3.9)
SP-E (Ours)	**100%**	**100%**	**93.6%**	**100%**	**100%**	**100%**	**100%**	**99.1%**	(2.3)
Reconstructed points ($\|3DPts\|$)									
SIFT	13470	6599	26225	5700	2505	7666	9608	10253.3	(7237.6)
SP	12941	9057	17451	6489	4093	8911	12535	10211.0	(4133.1)
SP-E (Ours)	**34851**	**45471**	**42727**	**33277**	**36403**	**19286**	**31851**	**34838.0**	(7846.5)
Mean reprojection error (Err)									
SIFT	**1.34**	**1.38**	**0.95**	**1.45**	**1.40**	**1.30**	**1.34**	**1.31**	(0.15)
SP	1.52	1.49	1.38	1.58	1.48	1.51	1.51	1.50	(0.06)
SP-E (Ours)	1.69	1.68	1.71	1.90	1.73	1.81	1.75	1.75	(0.07)
*Mean reprojection error of the best 10K points** (Err-10K)									
SIFT	1.08	1.38	**0.46**	1.45	1.40	**1.30**	1.34	1.20	(0.32)
SP	1.30	1.49	1.00	1.58	1.48	1.51	1.30	1.38	(0.19)
SP-E (Ours)	**0.92**	**0.73**	0.84	**1.30**	**0.91**	1.41	**1.06**	**1.02**	(0.23)
Mean track length (len(Tr))									
SIFT	6.57	5.73	**10.88**	12.48	7.74	**12.88**	7.56	9.12	(2.70)
SP	5.54	4.52	7.86	8.73	5.16	8.20	5.38	6.49	(1.59)
SP-E (Ours)	**7.05**	**6.78**	9.63	**14.73**	**8.42**	11.29	**8.78**	**9.53**	(2.55)

$^+$ Total number of images in the subsequence.
* If 10K points are not available, average is computed over all available reconstructed points.

the example in Fig. 2). The mean reprojection error of all the points is the lowest for SIFT, possibly due to it being more restrictive in all other aspects (number of images reconstructed, number of points, track length). However, the mean error for the top 10000 points is always lower for SP-E. The reprojection error plots in Figs. 3 and 4 provide more insight on this metric.

Figures 3 and 4 show a more detailed visualization of two representative reconstructions, including a summary of the sequence frames, the point cloud obtained by each method and a plot of the reprojection error for each point in the reconstruction, sorted in increasing error value. Note that even though SP-E obtains many more points, it is not at the cost of quality. Figure 3 shows a scenario where SIFT fails to reconstruct a large part of the subsequence, because it fails on the feature matching on the darker frames depicted in the middle of the sequence. Note how the reconstruction from SP-E is notably denser than the others. Figure 4 shows a scenario where all approaches perform well and SIFT achieves the lowest reprojection error.

Table 3. Analysis of the feature locations for each method.

	Spread of features ↑	% of features on specularities ↓
SIFT	43.9%	28.6%
SP	56.9%	19.6%
SP-E (Ours)	**67.5**%	**9.9**%

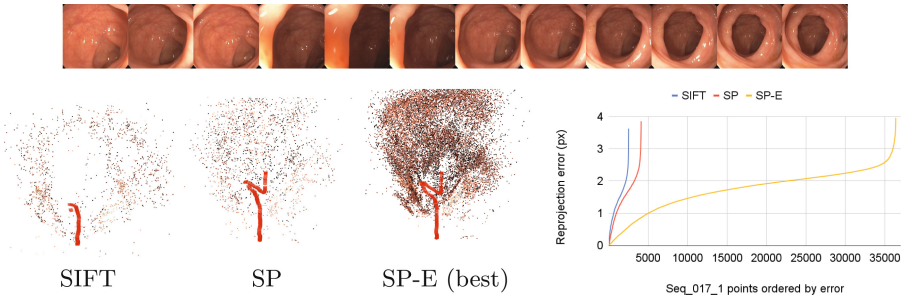

Fig. 3. Comparison of reconstructions obtained on Seq_017_1 by SIFT, SP, and our best model SP-E. The plot shows the reprojection error of each point reconstructed.

We analyze additional aspects of our detected features to showcase the higher quality with respect to other methods in Table 3. To measure the spread of the features over the images we defined a 16×16 grid over each image and computed the percentage of those cells that have at least one reconstructed point. We also measure how many extracted points fall on top of specularities (we consider a pixel as part of a specularity if its intensity is higher than 180). For both

metrics, our detector achieves significantly better results, showcasing the better properties of our detector for 3D reconstruction.

To provide quantitative evaluation of the camera motion estimation, we use a simulated dataset [3] to have ground truth available for the camera trajectory. We took 5 sequences of 100-150 frames from this dataset, and we tested the baselines and our model. We align the ground truth trajectories with the reconstructed ones with Horn's method [8]. SP only reconstructed 3 out of the 5 sequences while SIFT and SP-E correctly reconstructed the 5 sequences, with an average RMSE of 4.61mm and 4.71 mm respectively. Simulated data lacks some of the biggest challenges of endoscopy images (e.g. specularities, deformations), but this experiment suggests that the camera motion estimation quality is similarly good for all methods when they manage to converge.

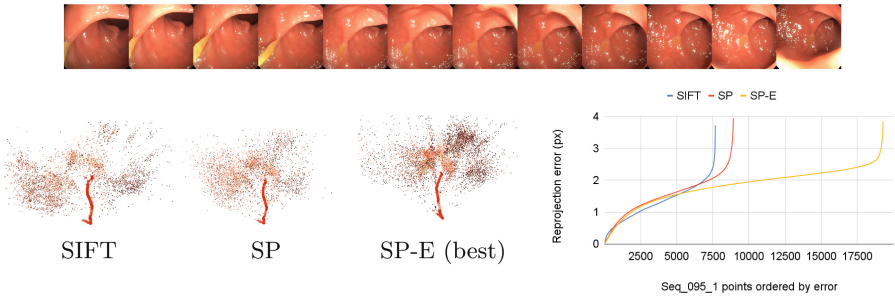

Fig. 4. Comparison of reconstructions obtained on Seq_095_1 by SIFT, SP, and our best model SP-E. The plot shows the reprojection error of each point reconstructed.

5 Conclusions

This work presents a novel training strategy for SuperPoint to improve its performance in SfM from endoscopy images. This strategy has two main benefits: we show how to use 3D reconstructions of endoscopy sequences as supervision to train feature extraction models; and we design a new tracking loss to perform *tracking adaptation* using this supervision. The benefits of our method are explored with an ablation study and against established baselines on SfM and feature extraction. Our proposed model is able to obtain more suitable features for 3D reconstruction, and to reconstruct larger sets of images with much denser point clouds.

Acknowledgements. This project has been funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No 863146 and Aragón Government project T45_23R.

References

1. Azagra, P., et al.: Endomapper dataset of complete calibrated endoscopy procedures. arXiv preprint arXiv:2204.14240 (2022)
2. Barbed, O.L., Chadebecq, F., Morlana, J., Montiel, J.M., Murillo, A.C.: Superpoint features in endoscopy. In: Imaging Systems for GI Endoscopy, and Graphs in Biomedical Image Analysis: First MICCAI Workshop, ISGIE 2022, and Fourth MICCAI Workshop, GRAIL 2022, Held in Conjunction with MICCAI 2022, Singapore, 18 September 2022, Proceedings, pp. 45–55. Springer, Heidelberg (2022). https://doi.org/10.1007/978-3-031-21083-9_5
3. Bobrow, T.L., Golhar, M., Vijayan, R., Akshintala, V.S., Garcia, J.R., Durr, N.J.: Colonoscopy 3d video dataset with paired depth from 2d–3d registration. arXiv preprint arXiv:2206.08903 (2022)
4. DeTone, D., Malisiewicz, T., Rabinovich, A.: Self-improving visual odometry. arXiv preprint arXiv:1812.03245 (2018)
5. DeTone, D., Malisiewicz, T., Rabinovich, A.: Superpoint: self-supervised interest point detection and description. In: Conference on Computer Vision and Pattern Recognition Workshops. IEEE (2018)
6. Dusmanu, M., et al.: D2-net: a trainable cnn for joint description and detection of local features. In: Conference on Computer Vision and Pattern Recognition. IEEE (2019)
7. Grasa, O.G., Bernal, E., Casado, S., Gil, I., Montiel, J.: Visual slam for handheld monocular endoscope. IEEE Trans. Med. Imaging **33**(1), 135–146 (2013)
8. Horn, B.K.: Closed-form solution of absolute orientation using unit quaternions. Josa a **4**(4), 629–642 (1987)
9. Jau, Y.Y., Zhu, R., Su, H., Chandraker, M.: Deep keypoint-based camera pose estimation with geometric constraints. In: International Conference on Intelligent Robots and Systems. IEEE (2020). https://github.com/eric-yyjau/pytorch-superpoint
10. Jin, Y., et al.: Image matching across wide baselines: from paper to practice. Int. J. Comput. Vision **129**(2), 517–547 (2021)
11. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision **60**(2), 91–110 (2004)
12. Ma, J., Jiang, X., Fan, A., Jiang, J., Yan, J.: Image matching from handcrafted to deep features: a survey. Int. J. Comput. Vision **129**, 1–57 (2020)
13. Mahmoud, N., Collins, T., Hostettler, A., Soler, L., Doignon, C., Montiel, J.M.M.: Live tracking and dense reconstruction for handheld monocular endoscopy. IEEE Trans. Med. Imaging **38**(1), 79–89 (2018)
14. Maier-Hein, L., et al.: Optical techniques for 3d surface reconstruction in computer-assisted laparoscopic surgery. Med. Image Anal. **17**(8), 974–996 (2013)
15. Mur-Artal, R., Montiel, J.M.M., Tardos, J.D.: Orb-slam: a versatile and accurate monocular slam system. IEEE Trans. Rob. **31**(5), 1147–1163 (2015)
16. Ozyoruk, K.B., et al.: Endoslam dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos. Med. Image Anal. **71**, 102058 (2021)
17. Revaud, J., Weinzaepfel, P., de Souza, C.R., Humenberger, M.: R2D2: repeatable and reliable detector and descriptor. In: International Conference on Neural Information Processing Systems (2019)
18. Rodríguez, J.J.G., Tardós, J.D.: Tracking monocular camera pose and deformation for slam inside the human body. In: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 5278–5285. IEEE (2022)

19. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: ORB: an efficient alternative to sift or surf. In: International Conference on Computer Vision. IEEE (2011)
20. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: Superglue: learning feature matching with graph neural networks. In: Conference on Computer Vision and Pattern Recognition. IEEE (2020)
21. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
22. Schönberger, J.L., Zheng, E., Frahm, J.-M., Pollefeys, M.: Pixelwise view selection for unstructured multi-view stereo. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9907, pp. 501–518. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46487-9_31
23. Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X.: Loftr: detector-free local feature matching with transformers. In: CVPR. IEEE (2021)
24. Tyszkiewicz, M., Fua, P., Trulls, E.: Disk: learning local features with policy gradient. Adv. Neural Inf. Process. Syst. **33**, 14254–14265 (2020)