Deep Homography Prediction for Endoscopic Camera Motion Imitation Learning

Martin Huber(✉), Sébastien Ourselin, Christos Bergeles, and Tom Vercauteren

School of Biomedical Engineering & Image Sciences, King's College London, London, UK

martin.huber@kcl.ac.uk

Abstract. In this work, we investigate laparoscopic camera motion automation through imitation learning from retrospective videos of laparoscopic interventions. A novel method is introduced that learns to augment a surgeon's behavior in image space through object motion invariant image registration via homographies. Contrary to existing approaches, no geometric assumptions are made and no depth information is necessary, enabling immediate translation to a robotic setup. Deviating from the dominant approach in the literature which consist of following a surgical tool, we do not handcraft the objective and no priors are imposed on the surgical scene, allowing the method to discover unbiased policies. In this new research field, significant improvements are demonstrated over two baselines on the Cholec80 and HeiChole datasets, showcasing an improvement of 47% over camera motion continuation. The method is further shown to indeed predict camera motion correctly on the public motion classification labels of the AutoLaparo dataset. All code is made accessible on GitHub (https://github.com/RViMLab/homography_imitation_learning).

Keywords: Computer vision · Robotic surgery · Imitation learning

1 Introduction

Automation in robot-assisted minimally invasive surgery (RMIS) may reduce human error that is linked to fatigue, lack of attention and cognitive overload [8]. It could help surgeons operate such systems by reducing the learning curve [29]. And in an ageing society with shrinking workforce, it could help to retain accessibility to healthcare. It is therefore expected that parts of RMIS will be ultimately automated [5,30]. On the continuous transition towards different levels of autonomy, camera motion automation is likely to happen first [14].

C. Bergeles and T. Vercauteren—These authors contributed equally to this work.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14228, pp. 217–226, 2023.
https://doi.org/10.1007/978-3-031-43996-4_21

Initial attempts to automate camera motion in RMIS include rule-based approaches that keep surgical tools in the center of the field of view [4,9,21]. The assumption that surgical tools remain centrally is, however, simplistic, as in many cases the surgeon may want to observe the surrounding anatomy to decide their next course of action.

Contrary to rule-based approaches, data-driven methods are capable to capture more complex control policies. Example data-driven methods suitable for camera motion automation include reinforcement learning (RL) and imitation learning (IL). The sample inefficiency and potential harm to the patient currently restrict RL approaches to simulation [1,22,23], where a domain gap remains. Work to bridge the domain gap and make RL algorithms deployable in real setups have been proposed [3,20], but clinical translation has not yet been achieved. For IL, on the other hand, camera motion automation could be learned from real data, thereby implicitly tackling the domain-gap challenge. The downside is that sufficient data may be difficult to collect. Many works highlight that lack of expert annotated data hinders progress towards camera motion automation in RMIS [7,13,19]. It is thus not surprising that existing literature on IL for camera motion automation utilizes data from mock setups [12,26].

Recent efforts to make vast amounts of laparoscopic intervention videos publicly available [19] drastically change how IL for camera motion automation can be approached. So far, this data is leveraged mainly to solve auxiliary tasks that could contribute to camera motion automation. As reviewed in [18], these tasks include tool and organ segmentation, as well as surgical phase recognition. For camera motion automation specifically, however, there exist no publicly available image-action pairs. Some work, therefore, continues to focus on the tools to infer camera motion [15], or learns on a robotic setup altogether [17] where camera motion is accessible. The realization, however, that camera motion is intrinsic to the videos of laparoscopic interventions and that camera motion could be learned on harvested actions was first realized in [11], and later in [16]. This comes with the additional advantage that no robot is necessary to learn behaviors and that one can directly learn from human demonstrations.

In this work, we build on [11] for computationally efficient image-action pair extraction from publicly available datasets of laparoscopic interventions, which yields more than $20\times$ the amount of data that was used in the closed source data of [16]. Contrary to [16], our camera motion extraction does not rely on image features, which are sparse in surgical videos, and is intrinsically capable to differentiate between camera and object motion. We further propose a novel importance sampling and data augmentation step for achieving camera motion automation IL.

2 Materials and Methods

The proposed approach to learning camera motion prediction is summarized in Fig. 1. The following sections will describe its key components in more detail.

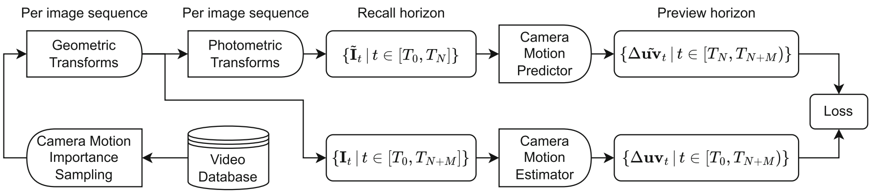


Fig. 1. Training pipeline, refer to Sect. 2.3. From left to right: Image sequences are importance sampled from the video database and random augmentations are applied per sequence online. The lower branch estimates camera motion between subsequent frames, which is taken as pseudo-ground-truth for the upper branch, which learns to predict camera motion on a preview horizon.

2.1 Theoretical Background

Points on a plane, as observed from a moving camera, transform by means of the 3×3 projective homography matrix $\mathbf{G}$ in image space. Thus, predicting future camera motion (up to scale) may be equivalently treated as predicting future projective homographies.

It has been shown in [6] that the four point representation of the projective homography, *i.e.,* taking the difference between four points in homogeneous coordinates $\Delta\mathbf{uv} = \{\mathbf{p}_i - \mathbf{p}'_i \,|\, i \in [0, 4)\} \in \mathbb{R}^{4\times 2}$ that are related by $\mathbf{G}\mathbf{p}_i \sim \mathbf{p}'_i \;\forall i$, is better behaved for deep learning applications than the 3×3 matrix representation of a homography. Therefore, in this work, we treat camera motion $\mathcal{C}$ as a sequence of four point homographies on a time horizon $[T_0, T_{N+M})$, N being the recall horizon's length, M being the preview horizon's length. Time points lie Δt apart, that is $T_{i+1} = T_i + \Delta t$. For image sequences of length N+M, we work with four point homography sequences $\mathcal{C} = \{\Delta\mathbf{uv}_t \,|\, t \in [T_0, T_{N+M})\}$.

2.2 Data and Data Preparation

Three datasets are curated to train and evaluate the proposed method: two cholecystectomy datasets (laparoscopic gallbladder removal), namely Cholec80 [25] and HeiChole [27], and one hysterectomy dataset (laparoscopic uterus removal), namely AutoLaparo [28].

To remove status indicator overlays from the laparoscopic videos, which may hinder the camera motion estimator, we identify the bounding circle of the circular field of view using [2]. We crop the view about the center point of the bounding circle to a shape of 240×320, so that no black regions are prominent in the images.

All three datasets are split into training, validation, and testing datasets. We split the videos by frame count into $80 \pm 1\%$ training and $20 \pm 1\%$ testing. Training and testing videos never intersect. We repeat this step to further split the training dataset into (pure) training and validation datasets.

Due to errors during processing the raw data, we exclude videos 19, 21, and 23 from HeiChole, as well as videos 22, 40, 65, and 80 from Cholec80. This results in dataset sizes of: Cholec80 - $4.4e6$ frames at 25 fps, HeiChole - $9.5e5$ frames at 25 fps, and AutoLaparo - $7.1e4$ frames at 25 fps.

2.3 Proposed Pipeline

Video Database and Importance Sampling. The curated data from Sect. 2.2 is accumulated into a video database. Image sequences of length $N+M$ are sampled at a frame increment of Δn between subsequent frames and with Δc frames between the sequence's initial frames. Prior to adding the videos to the database, an initial offline run is performed to estimate camera motion $\Delta \mathbf{uv}$ between the frames. This creates image-motion correspondences of the form $(\mathbf{I}_n, \mathbf{I}_{n+\Delta n}, \Delta \mathbf{uv}_n)$. Image-motion correspondences where $\mathbb{E}(||\Delta \mathbf{uv}_n||_2) > \sigma$, with sigma being the standard deviation over all motions in the respective dataset, define anchor indices n. Image sequences are sampled such that the last image in the recall horizon lies at index $n = N - 1$, marking the start of a motion. The importance sampling samples indices from the intersection of all anchor indices, shifted by $-N$, with all possible starting indices for image sequences.

Geometric and Photometric Transforms. The importance sampled image sequences are fed to a data augmentation stage. This stage entails geometric and photometric transforms. The distinction is made because downstream, the pipeline is split into two branches. The upper branch serves as camera motion prediction whereas the lower branch serves as camera motion estimation, also refer to the next section. As it acts as the source of pseudo-ground-truth, it is crucial that the camera motion estimator performs under optimal conditions, hence no photometric transforms, i.e. transforms that change brightness/contrast/fog etc., are applied. Photometrically transformed images shall further be denoted as $\tilde{\mathbf{I}}$. To encourage same behavior under different perspectives, geometric transforms are applied, i.e. transforms that change orientation/up to down/left to right etc. Transforms are always sampled randomly, and applied consistently to the entire image sequence.

Camera Motion Estimator and Predictor. The goal of this work is to have a predictor learn camera motion computed by an estimator. The predictor takes as input a photometrically and geometrically transformed recall horizon $\{\tilde{\mathbf{I}}_t \,|\, t \in [T_0, T_N)\}$ of length N, and predicts camera motion $\tilde{\mathcal{C}} = \{\Delta \tilde{\mathbf{uv}}_t \,|\, t \in [T_N, T_{N+M})\}$ on the preview horizon of length M. The estimator takes as input the geometrically transformed preview horizon $\{\mathbf{I}_t \,|\, t \in [T_M, T_{N+M})\}$ and estimates camera motion $\mathcal{C}$, which serves as a target to the predictor. The estimator is part of the pipeline to facilitate on-the-fly perspective augmentation via the geometric transforms.

3 Experiments and Evaluation Methodology

The following two sections elaborate the experiments we conduct to investigate the proposed pipeline from Fig. 1 in Sect. 2.3. First the camera motion estimator is investigated, followed by the camera motion predictor.

3.1 Camera Motion Estimator

Camera Motion Distribution. To extract the camera motion distribution, we run the camera motion estimator from [11] with a ResNet-34 backbone over all datasets from Sect. 2.2. We map the estimated four point homographies to up/down/left/right/zoom-in/zoom-out for interpretability. Left/right/up/down corresponds to all four point displacements $\Delta\mathbf{uv}$ consistently pointing left/right/up/down respectively. Zoom-in/out corresponds to all four point displacements $\Delta\mathbf{uv}$ consistently pointing inwards/outwards. Rotation left corresponds to all four point displacements pointing up right, bottom right, and so on. Same for rotation right. Camera motion is defined static if it lies below the standard deviation in the dataset. The frame increment is set to 0.25 s, corresponding to $\Delta n = 5$ for the 25 fps videos.

Online Camera Motion Estimation. Since the camera motion estimator is executed online, memory footprint and computational efficiency are of importance. Therefore, we evaluate the estimator from [11] with a ResNet-34 backbone, SURF & RANSAC, and LoFTR [24] & RANSAC. Each estimator is run 1000 times on a single image sequence of length $N+M = 15$ with an NVIDIA GeForce RTX 2070 GPU and an Intel(R) Core(TM) i7-9750H CPU @ 2.60 GHz.

3.2 Camera Motion Predictor

Model Architecture. For all experiments, the camera motion predictor is a ResNet-18/34/50, with the number of input features equal to the recall horizon $\mathrm{N} \times 3$ (RGB), where $\mathrm{N} = 14$. We set the preview horizon $\mathrm{M} = 1$. The frame increment is set to 0.25 s, or $\Delta n = 5$ for the 25 fps videos. The number of frames between clips is also set to 0.25 s, or $\Delta c = 5$.

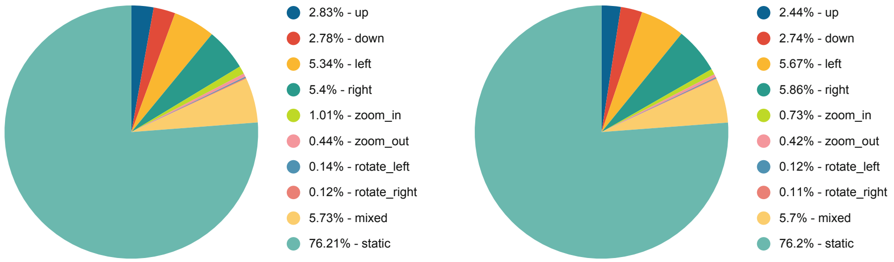
Training Details. The camera motion predictor is trained on each dataset from Sect. 2.2 individually. For training on Cholec80/HeiChole/AutoLaparo, we run 80/50/50 epochs on a batch size of 64 with a learning rate of $2.5e-5/1.e-4/1.e-4$. The learning rates for Cholec80 and HeiChole relate approximately to the dataset's training sizes, see Table 2. For Cholec80, we reduce the learning rate by a factor 0.5 at epochs 50, 75. For Heichole/AutoLaparo we drop the learning rate by a factor 0.5 at epoch 35. The loss in Fig. 1 is set to the mean pairwise distance between estimation and prediction $\mathbb{E}(||\Delta\tilde{\mathbf{uv}}_t - \Delta\mathbf{uv}_t||_2) + \lambda\mathbb{E}(||\Delta\tilde{\mathbf{uv}}_t||_2)$ with a regularizer that discourages the identity $\Delta\tilde{\mathbf{uv}}_t = \mathbf{0}$ (i.e. no motion). We set $\lambda = 0.1$.

Evaluation Metrics. For evaluation we compute the mean pairwise distance between estimated and predicted motion $\mathbb{E}(||\Delta\tilde{\mathbf{uv}}_t - \Delta\mathbf{uv}_t||_2)$. All camera motion predictors are benchmarked against a baseline, that is a $\mathcal{O}(1)/\mathcal{O}(2)$-Taylor expansion of the estimated camera motion $\Delta\mathbf{uv}_t$. Furthermore, the model that is found to perform best is evaluated on the multi-class labels (left, right, up, down) that are provided in AutoLaparo.

4 Results

4.1 Camera Motion Estimator

Camera Motion Distribution. The camera motion distributions for all datasets are shown in Fig. 2. It is observed that for a large fraction of the sequences there is no significant camera motion (Cholec80 76.21%, HeiChole 76.2%, AutoLaparo 71.29%). This finding supports the importance sampling that was introduced in Sect. 2.3. It can further be seen that e.g. left/right and up/down motions are equally distributed.



(a) Cholec80 (b) HeiChole

Fig. 2. Camera motion distribution, refer to Sect. 3.1. AutoLaparo: 2.81% - up, 1.88% - down, 4.48% - left, 3.38% - right, 0.45% - zoom_in, 0.2% - zoom_out, 0.3% - rotate_left 0.3%, - rotate_right 14.9% - mixed, 71.29% - static.

Online Camera Motion Estimation. The results of the online camera motion estimation are summarized in Table 1. The deep homography estimation with a Resnet34 backbone executes 11× quicker and has the lowest GPU memory footprint of the GPU accelerated methods. This allows for efficient implementation of the proposed online camera motion estimation in Fig. 1.

4.2 Camera Motion Prediction

The camera motion prediction results for all datasets are highlighted in Table 2. It can be seen that significant improvements over the baseline are achieved on the Cholec80 and HeiChole datasets. Whilst the learned prediction performs better on average than the baseline, no significant improvement is found for the AutoLaparo dataset.

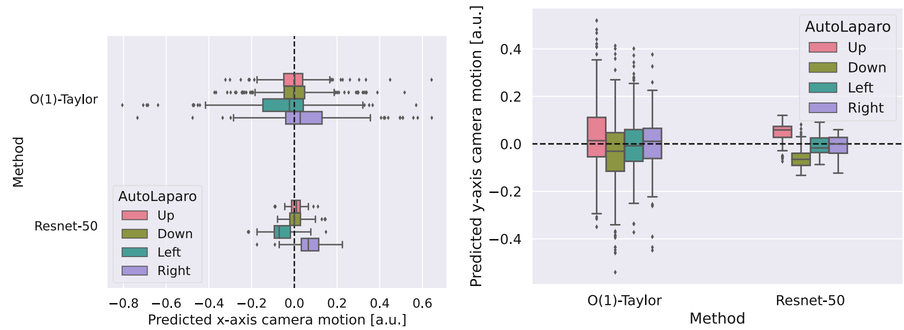
Table 1. Memory footprint and execution time of different camera motion estimators, refer to Sect. 3.1.

Method	Execution time [s]	Speed-up [a.u.]	Model/Batch [Mb]
Resnet34	$\mathbf{0.016 \pm 0.048}$	**11.1**	**664/457**
LoFTR & RANSAC	0.178 ± 0.06	1.0	669/2412
SURF & RANSAC	0.131 ± 0.024	1.4	NA

The displacement of the image center point under the predicted camera motion for AutoLaparo is plotted against the provided multi-class motion annotations and shown in Fig. 3. It can be seen that the camera motion predictions align well with the ground truth labels.

Table 2. Camera motion predictor performance, refer to Sect. 3.2. Taylor baselines predict based on previous estimated motion, ResNets based on images.

Dataset	Train Size [Frames]	Mean Pairwise Distance [Pixels]				
		Taylor		ResNet (proposed)		
		$\mathcal{O}(1)$	$\mathcal{O}(2)$	18	34	50
Cholec80	$3.5e6$	27.2 ± 23.1	36.4 ± 31.2	$\mathbf{14.8} \pm 11.7$	$\mathbf{14.4} \pm 11.4$	$\mathbf{14.4} \pm 11.4$
HeiChole	$7.6e5$	29.7 ± 26.4	39.8 ± 35.9	$\mathbf{15.8} \pm 12.5$	$\mathbf{15.8} \pm 12.5$	$\mathbf{15.8} \pm 12.5$
AutoLaparo	$5.9e4$	19.4 ± 18.4	25.8 ± 24.7	$\mathbf{11.2} \pm 11.0$	$\mathbf{11.3} \pm 11.0$	$\mathbf{11.3} \pm 11.0$



(a) Predicted camera motion along x-axis, scaled by image size to $[-1, 1]$.

(b) Predicted camera motion along y-axis, scaled by image size to $[-1, 1]$.

Fig. 3. Predicted camera motion on AutoLaparo, refer to Sect. 3.2. Camera motion predictor trained on Cholec80 with ResNet-50 backbone, see Table 2. Shown is the motion of the image center under the predicted homography. Clearly, for videos labeled left/right, the center point is predicted to move left/right and for up/down labels, the predicted left/right motion is centered around zero (a). Same is observed for up/down motion in (b), where left/right motion is zero-centered.

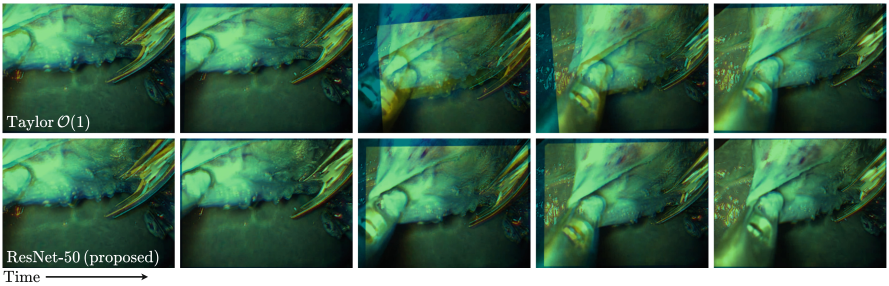


Fig. 4. Exemplary camera motion prediction, refer to Sect. 3.2. In the image sequence, the attention changes from the right to the left tool. We warp the past view (yellow) by the predicted homography and overlay the current view (blue). Good alignment corresponds to good camera motion prediction. Contrary to the baseline, the proposed method predicts the motion well. Data taken from HeiChole test set, ResNet-50 backbone trained on Cholec80, refer Table 2. (Color figure online)

5 Conclusion and Outlook

To the best of our knowledge, this work is the first to demonstrate that camera motion can indeed be learned from retrospective videos of laparoscopic interventions, with no manual annotation. Self-supervision is achieved by harvesting image-motion correspondences using a camera motion estimator, see Fig. 1. The camera motion predictor is shown to generate statistically significant better predictions over a baseline in Table 2 as measured using pseudo-ground-truth and on multi-class manually annotated motion labels from AutoLaparo in Fig. 3. An exemplary image sequence in Fig. 4 demonstrates successful camera motion prediction on HeiChole. These results were achieved through the key finding from Fig. 2, which states that most image sequences, i.e. static ones, are irrelevant to learning camera motion. Consequentially, we contribute a novel importance sampling method, as described in Sect. 2.3. Finally, we hope that our open-source commitment will help the community explore this area of research further.

A current limitations of this work is the preview horizon M of length 1. One might want to extend it for model predictive control. Furthermore, to improve explainability to the surgeon, but also to improve the prediction in general, it would be beneficial to include auxiliary tasks, e.g. tool and organ segmentation, surgical phase recognition, and audio. There also exist limitations for the camera motion estimator. The utilized camera motion estimator is efficient and isolates object motion well from camera motion, but is limited to relatively small camera motions. Improving the camera motion estimator to large camera motions would help increase the preview horizon M.

In future work, we will execute this model in a real setup for investigating transferability. This endeavor is backed by [10], which demonstrates how the learned homography could immediately be deployed on a robotic laparoscope holder. It might proof necessary to fine-tune the presented policy through reinforcement learning with human feedback.

Acknowledgements. This work was supported by core and project funding from the Wellcome/EPSRC [WT203148/Z/16/Z; NS/A000049/1; WT101957; NS/A000027/1]. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016985 (FAROS project). TV is supported by a Medtronic/RAEng Research Chair [RCSRF1819\7\34]. SO and TV are co-founders and shareholders of Hypervision Surgical. TV is co-founder and shareholder of Hypervision Surgical. TV holds shares from Mauna Kea Technologies.

References

1. Agrawal, A.S.: Automating endoscopic camera motion for teleoperated minimally invasive surgery using inverse reinforcement learning. Ph.D. thesis, Worcester Polytechnic Institute (2018)
2. Budd, C., Garcia-Peraza Herrera, L.C., Huber, M., Ourselin, S., Vercauteren, T.: Rapid and robust endoscopic content area estimation: a lean GPU-based pipeline and curated benchmark dataset. Comput. Methods Biomech. Biomed. Eng. Imaging Vis. **11**(4), 1215–1224 (2022). https://doi.org/10.1080/21681163.2022.2156393
3. Cartucho, J., Tukra, S., Li, Y., Elson, D.S., Giannarou, S.: VisionBlender: a tool to efficiently generate computer vision datasets for robotic surgery. Comput. Methods Biomech. Biomed. Eng. Imaging Vis. **9**(4), 331–338 (2021)
4. Da Col, T., Mariani, A., Deguet, A., Menciassi, A., Kazanzides, P., De Momi, E.: SCAN: system for camera autonomous navigation in robotic-assisted surgery. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 2996–3002. IEEE (2020)
5. Davenport, T., Kalakota, R.: The potential for artificial intelligence in healthcare. Future Healthc. J. **6**(2), 94 (2019)
6. DeTone, D., Malisiewicz, T., Rabinovich, A.: Deep image homography estimation (2016). http://arxiv.org/abs/1606.03798
7. Esteva, A., et al.: A guide to deep learning in healthcare. Nat. Med. **25**(1), 24–29 (2019)
8. Fiorini, P., Goldberg, K.Y., Liu, Y., Taylor, R.H.: Concepts and trends in autonomy for robot-assisted surgery. Proc. IEEE **110**(7), 993–1011 (2022)
9. Garcia-Peraza-Herrera, L.C., et al.: Robotic endoscope control via autonomous instrument tracking. Front. Robot. AI **9**, 832208 (2022)
10. Huber, M., Mitchell, J.B., Henry, R., Ourselin, S., Vercauteren, T., Bergeles, C.: Homography-based visual servoing with remote center of motion for semi-autonomous robotic endoscope manipulation. In: 2021 International Symposium on Medical Robotics (ISMR), pp. 1–7. IEEE (2021)
11. Huber, M., Ourselin, S., Bergeles, C., Vercauteren, T.: Deep homography estimation in dynamic surgical scenes for laparoscopic camera motion extraction. Comput. Methods Biomech. Biomed. Eng. Imaging Visu. **10**(3), 321–329 (2022)
12. Ji, J.J., Krishnan, S., Patel, V., Fer, D., Goldberg, K.: Learning 2D surgical camera motion from demonstrations. In: 2018 IEEE 14th International Conference on Automation Science and Engineering (CASE), pp. 35–42. IEEE (2018)
13. Kassahun, Y., et al.: Surgical robotics beyond enhanced dexterity instrumentation: a survey of machine learning techniques and their role in intelligent and autonomous surgical actions. Int. J. Comput. Assist. Radiol. Surg. **11**, 553–568 (2016). https://doi.org/10.1007/s11548-015-1305-z

14. Kitaguchi, D., Takeshita, N., Hasegawa, H., Ito, M.: Artificial intelligence-based computer vision in surgery: recent advances and future perspectives. Ann. Gastroenterological Surg. **6**(1), 29–36 (2022)
15. Li, B., Lu, B., Lu, Y., Dou, Q., Liu, Y.H.: Data-driven holistic framework for automated laparoscope optimal view control with learning-based depth perception. In: 2021 IEEE International Conference on Robotics and Automation (ICRA), pp. 12366–12372. IEEE (2021)
16. Li, B., Lu, B., Wang, Z., Zhong, F., Dou, Q., Liu, Y.H.: Learning laparoscope actions via video features for proactive robotic field-of-view control. IEEE Robot. Autom. Lett. **7**(3), 6653–6660 (2022)
17. Li, B., et al.: 3D perception based imitation learning under limited demonstration for laparoscope control in robotic surgery. In: 2022 International Conference on Robotics and Automation (ICRA), pp. 7664–7670. IEEE (2022)
18. Loukas, C.: Video content analysis of surgical procedures. Surg. Endosc. **32**, 553–568 (2018). https://doi.org/10.1007/s00464-017-5878-1
19. Maier-Hein, L., et al.: Surgical data science-from concepts toward clinical translation. Med. Image Anal. **76**, 102306 (2022)
20. Marzullo, A., Moccia, S., Catellani, M., Calimeri, F., De Momi, E.: Towards realistic laparoscopic image generation using image-domain translation. Comput. Methods Programs Biomed. **200**, 105834 (2021)
21. Sandoval, J., Laribi, M.A., Faure, J., Breque, C., Richer, J.P., Zeghloul, S.: Towards an autonomous robot-assistant for laparoscopy using exteroceptive sensors: feasibility study and implementation. IEEE Robot. Autom. Lett. **6**(4), 6473–6480 (2021)
22. Scheikl, P.M., et al.: LapGym-an open source framework for reinforcement learning in robot-assisted laparoscopic surgery. arXiv preprint arXiv:2302.09606 (2023)
23. Su, Y.H., Huang, K., Hannaford, B.: Multicamera 3D viewpoint adjustment for robotic surgery via deep reinforcement learning. J. Med. Robot. Res. **6**(01n02), 2140003 (2021)
24. Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X.: LoFTR: detector-free local feature matching with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8922–8931 (2021)
25. Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., De Mathelin, M., Padoy, N.: EndoNet: a deep architecture for recognition tasks on laparoscopic videos. IEEE Trans. Med. Imaging **36**(1), 86–97 (2016)
26. Wagner, M., et al.: A learning robot for cognitive camera control in minimally invasive surgery. Surg. Endosc. **35**(9), 5365–5374 (2021). https://doi.org/10.1007/s00464-021-08509-8
27. Wagner, M., et al.: Comparative validation of machine learning algorithms for surgical workflow and skill analysis with the heichole benchmark. Med. Image Anal. **86**, 102770 (2023)
28. Wang, Z., et al.: AutoLaparo: a new dataset of integrated multi-tasks for image-guided surgical automation in laparoscopic hysterectomy. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022, Part VII. LNCS, vol. 13437, pp. 486–496. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16449-1_46
29. van Workum, F., Fransen, L., Luyer, M.D., Rosman, C.: Learning curves in minimally invasive esophagectomy. World J. Gastroenterol. **24**(44), 4974 (2018)
30. Zidane, I.F., Khattab, Y., Rezeka, S., El-Habrouk, M.: Robotics in laparoscopic surgery-a review. Robotica **41**(1), 126–173 (2023)