

Jumpstarting Surgical Computer Vision

Deepak Alapatt¹, Aditya Murali¹, Vinkle Srivastav^{1,2}, AI4SafeChole Consortium^{3,4,5,6}, Pietro Mascagni^{2,3}, and Nicolas Padoy^{1,2}

¹ University of Strasbourg, CNRS, INSERM, ICube, UMR7357, Strasbourg, France

² IHU-Strasbourg, Institute of Image-Guided Surgery, Strasbourg, France

³ Fondazione Policlinico Universitario Agostino Gemelli IRCCS, Rome, Italy

⁴ Azienda Ospedaliero-Universitaria Sant'Andrea, Rome, Italy

⁵ Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico di Milano, University of Milan, Milan, Italy

⁶ Monaldi Hospital, AORN dei Colli, Naples, Italy

Abstract. Consensus amongst researchers and industry points to a lack of large, representative annotated datasets as the biggest obstacle to progress in the field of surgical data science. Advances in Self-Supervised Learning (SSL) represent a solution, reducing the dependence on large labeled datasets by providing task-agnostic initializations. However, the robustness of current self-supervised learning methods to domain shifts remains unclear, limiting our understanding of its utility for leveraging diverse sources of surgical data. Shifting the focus from methods to data, we demonstrate that the downstream value of SSL-based initializations is intricately intertwined with the composition of pre-training datasets. These results underscore an important gap that needs to be filled as we scale self-supervised approaches toward building general-purpose “foundation models” that enable diverse use-cases within the surgical domain. Through several stages of controlled experimentation, we develop recommendations for pretraining dataset composition evidenced through over 300 experiments spanning 20 pre-training datasets, 9 surgical procedures, 7 centers (hospitals), 3 labeled-data settings, 3 downstream tasks, and multiple runs. Using the approaches here described, we outperform state-of-the-art pre-trainings on two public benchmarks for phase recognition: up to 2.2% on Cholec80 and 5.1% on AutoLaparo.

Keywords: Self-supervised learning · Surgical computer vision · Endoscopic videos · Critical View of Safety · Surgical phase recognition.

1 Introduction

In the domain of surgery, the potential of artificial intelligence to disrupt clinical practice is becoming increasingly apparent with the development of models that can democratize expert-level decision-making [14, 9], provide real-time intraoperative support [12], and generate video documentation [13] of procedures among various other applications. Still, a recent survey conducted by the Surgical Data Science Initiative [11], attributed a lack of tangible success stories to a lack

of representative annotated data [10]. Self-supervised learning (SSL) presents a promising approach to mitigating this reliance on large, well-annotated datasets by learning generic representations, and consequently, building generic model initializations. If sufficiently rich in information, these initializations could enable the development of foundation models for surgery, task-agnostic models that could robustly enable a diverse range of tasks with little to no labels.

SSL methodology in the broader computer vision domain has seen significant developments in design, moving from heuristic-based pre-training tasks, such as predicting image augmentations, to more elegant strategies that directly learn to predict latent representations specific to the content of an image. Until recently, these shifts in design were slow to be adopted into the surgical domain with most work focusing on hand-crafted pre-text tasks, such as colorization [18], or leveraging intrinsically collected information such as robot kinematics [3, 19] or the temporal order of frames [5, 16]. A large benchmark study of state-of-the-art SSL methodologies applied to the surgical domain attributed this to the complexity of translating and tuning newly proposed models to the surgical domain [17]. However, in the same study, the authors demonstrate through extensive experimentation that when applied correctly, these methods achieve state-of-the-art results on a range of tasks, benchmarks, and labeled-data regimes. Since then, several works have explored the value of pre-training similar methods on a massive scale by aggregating public datasets [1], utilizing large private repositories [6], or some combination of the two [22]. All three of these works demonstrate that large-scale pre-training on their respective datasets brings sizable boosts in performance when performing various downstream tasks. [1] aggregate 10 public datasets for several types of laparoscopic procedures sourced from centers worldwide to create a pre-training dataset of about 700k images. [6] compile a private dataset of ~ 7900 laparoscopic procedures ($> 23\text{M}$ images) from 8 centers and a curated polyp image dataset generated from $\sim 14\text{k}$ colonoscopies ($> 2\text{M}$ images). [22] combine 6 public datasets for laparoscopy and endoscopy along with a large private repository of endoscopic videos to pre-train on $\sim 33\text{k}$ videos and $\sim 5\text{M}$ images. Different from these works that aggregate laparoscopic and endoscopic video data collected from different centers (hospitals) and different procedures to pre-train at scale, we focus this study on methodically clarifying how surgical pre-training datasets could and should be scaled to optimally leverage diverse surgical data.

Our contributions are threefold:

1. We highlight important limitations when applying existing methodology, outperforming state-of-the-art approaches on two well-established public benchmarks in surgical data science, Cholec80 and AutoLaparo.
2. We quantify the sensitivity of pre-training efficacy to various factors in over 300 experiments spanning 20 pre-training datasets, 9 procedures, 7 centers (hospitals), 3 labeled-data settings, 3 downstream tasks, and multiple runs.
3. We validate our findings at scale (~ 400 videos) exploring different strategies to scale surgical SSL pre-training, quantifying key factors when doing so.

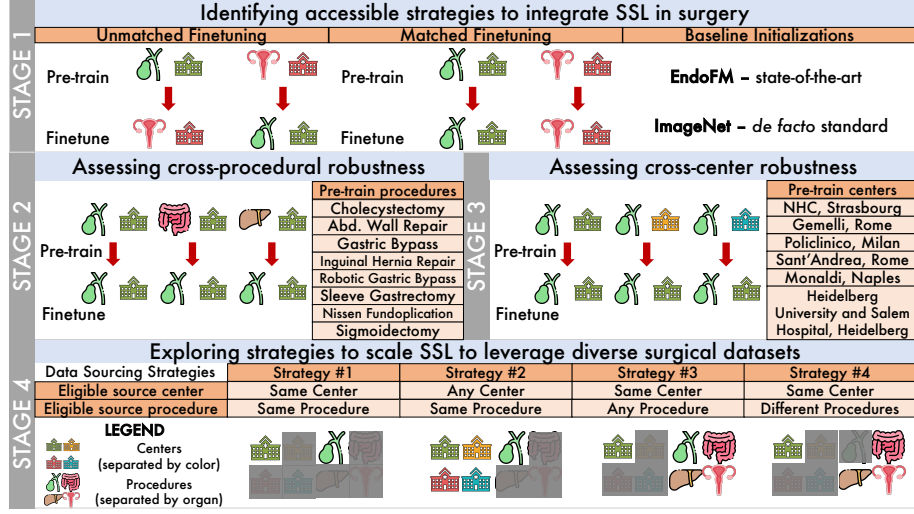


Fig. 1. Overview of experimental design, aimed at addressing different objectives through several stages of study and over 300 experiments.

2 Methodology

In this section, we break down the various datasets and downstream tasks used to enable several distinct study objectives summarized in Figure 1.

2.1 Datasets

We use 7 different surgical computer vision datasets of minimally invasive abdominal procedures including: AutoLaparo [23], Cholec80 [20], Endoscapes-CVS201 [15], HeiChole [21], and Laparo425 [7]. We also use two private unlabeled Laparoscopic Cholecystectomy (LC) datasets: MultiChole2024 and StrasChole400, each containing ~ 400 randomly selected cholecystectomy videos. StrasChole400 contains 400 videos from Strasbourg while MultiChole2024, an extension of MultiChole2022 [8], contains 405 videos, 200 from Strasbourg and ~ 50 videos from each of the 4 centers (listed in Figure 1).

These datasets span 10 different surgical procedures (several cholecystectomy datasets, 7 others in Laparo420, 1 other in AutoLaparo), and 7 centers (several Strasbourg datasets, 4 others in MultiChole2024, 2 aggregated in HeiChole). Lastly, for fair experimentation, we remove 5 videos from Laparo425 that are found in the test sets for the downstream tasks, yielding Laparo420. Using these datasets, we create our various pre-training data combinations.

2.2 Downstream Tasks

We consider 3 downstream tasks, each using a different dataset: phase recognition (on both Cholec80 and AutoLaparo) and critical view of safety (CVS)

prediction (Endoscopes-CVS201), thus spanning 2 procedures (cholecystectomy, hysterectomy) and 2 centers (Strasbourg, Hong Kong). Our selected downstream tasks are also highly diverse: CVS criteria assessment (3-class binary classification) is one of several tasks proposed for the prevention of bile duct injuries, requiring fine-grained analysis. Meanwhile, phase recognition is a widely studied and benchmarked task in surgical computer vision that, unlike CVS classification, requires long-range temporal reasoning. For each of the 2 phase recognition datasets, we consider one high-label data regime using 100% of the available training labels (40 videos for Cholec80, 10 videos for Autolaparo) and two low-label data regimes using 5 videos and 3 videos, respectively. For the CVS task, we similarly consider 3 label settings using 120 videos (100% of the train set), 30 videos (25% of the train set), and 15 videos (12.5% of the train set). For every low-label setting on each downstream task, we present results using the mean and standard deviation over 3 randomly sub-selected splits to mitigate the effects of selection bias. Following the bulk of previous works, we use F1-score to evaluate the phase recognition tasks and mean Average Precision (mAP) for CVS assessment.

2.3 Training Procedure

For all experiments, we use MoCo v2 [2] to pre-train a ResNet-50 feature extractor. MoCo v2 is a contrastive learning method that learns powerful image representations by minimizing differences between embeddings of augmented views of the same image to be similar and maximizing distance between embeddings of different images. We select MoCo v2 for all pre-training motivated by the results in [17], where it outperforms several competing methods across various surgical video analysis benchmarks. This pre-training procedure results in a trained ResNet-50 backbone, which we use to initialize a ResNet-50 classifier that we finetune separately for each downstream task while varying the number of labeled videos. Finally, for the phase recognition downstream tasks (Cholec80, AutoLaparo), we further train TeCNO [4] on top of the finetuned ResNet-50 features to enable temporal reasoning as in [17]. Detailed implementation settings are listed in the supplementary material.

2.4 Study Stages

Stage 1. The purpose of this stage is to establish performance baselines, exploring different application strategies of current SSL methodology to leverage related datasets (e.g. laparoscopic videos of related surgical procedures) for various tasks. To do this we perform 4 cross-over experiments examining the added value over ImageNet initializations: (1) where the downstream task is from the same dataset as the pre-training data (matched finetuning), and (2) where the downstream task and dataset are different from the pre-training dataset (unmatched finetuning). Note that both downstream tasks are conceptually similar (surgical phase recognition); however, the two datasets span two geographic regions (Hong Kong and France), two anatomical targets (Gallbladder and Uterus), and likely

stylistic and functional differences in workflow and instrumentation. We adopt this process for two datasets, Cholec80 and AutoLaparo, reporting both matched and unmatched finetuning performance for both downstream phase recognition tasks. We only utilize public datasets in this stage to additionally illustrate the potential of utilizing accessible datasets. We also compare these strategies to EndoFM [22], the state-of-the-art massive-scale pretraining of a ViT-B model aimed at creating a foundation model for endoscopic data.

Stage 2. In this stage, we aim to study the impact of pre-training surgical procedure types on downstream task performance. To enable this analysis, we separate Laparo420, which is a collection of various surgical videos, into 8 different subsets by procedure type (listed in Figure 1). We then evaluate on Cholec80 for phase recognition and Endoscapes-CVS201 for CVS Classification. In this controlled experiment, we keep the source center for all the pre-training datasets fixed (Strasbourg) and the pre-training dataset size fixed to ~ 50 videos⁷; by controlling the source center and pre-training data scale in this fashion, we aim to isolate the impact of the pre-training procedure type. Still, confounding variables such as differences in instrumentation, workflow etc. do exist, which we aim to mitigate by including a large number of procedures (8) and large number of cases per procedure (~ 50).

Stage 3. In the third stage of our study, we aim to study the impact of the source center for the pre-training dataset on downstream task performance. As has been noted by numerous works, surgical videos can vary greatly among medical centers due to differences in instrumentation, workflow, lighting, acquisition hardware, software, and patient characteristics. We leverage 4 subsets of MultiChole2024, which contains laparoscopic cholecystectomy videos from four different Italian hospitals, as well as the HeiChole public train set, combining cholecystectomy videos from 2 German hospitals. Further, we also utilize a 25-video subset of Cholec80; thus, constructing 6 subsets of ~ 25 videos. As in stage 2, this allows us to control for procedure type (cholecystectomy) and data scale (~ 25 videos)⁸ and isolate the impact of source center. Note again, confounding variables such as differences due to instrumentation, workflow, etc. do exist, which we aim to mitigate by including a large number of centers (2 from HeiChole and 5 others) and a large number of cases per center (~ 25). We again evaluate phase recognition and CVS classification.

Stage 4. Lastly, one of the great promises of SSL is its scalability; however, there are numerous possibilities to accomplish this scaling in the surgical domain. To name a few, only including data from the same center and procedure type as the downstream task, fixing the center and varying procedure type or vice-versa, or naively including any and all available videos. To identify an optimal scaling strategy, we pre-train models on a series of data combinations, enumerated in

⁷ All stage 2 datasets contain exactly 50 videos except for cholecystectomy (45 videos), as 5 videos were removed to avoid contamination with the downstream task test sets.

⁸ All center-specific pre-training datasets contain exactly 25 videos except for the public HeiChole training dataset, which contains 24 videos.

Pre-training Dataset	# Videos	Source Center		Procedure Type	
		Strasbourg	Other	Cholecystectomy	Other
StrasChole400	400	400	0	400	0
MultiChole2024	405	200	205	405	0
Laparo420 (with LC)	420	420	0	45	375
Laparo420 (w/o LC)	375	375	0	0	375

Table 1. Large Scale Pre-training Data Combinations

Initialization	Arch.	Cholec80 Phase (F1)			AutoLaparo Phase (F1)		
		Number of Labeled Videos					
		40	5	3	10	5	3
Cholec80 SSL (Ours)	R50	79.6	66.1 $\pm$ 4.2	61.7 $\pm$ 4.8	61.5	58.3 $\pm$ 2.8	47.7 $\pm$ 2.3
AutoLaparo SSL (Ours)	R50	80.0	62.4 $\pm$ 6.0	54.4 $\pm$ 5.7	70.7	64.4 $\pm$ 5.1	59.5 $\pm$ 3.9
EndoFM	ViT-B	78.4	65.9 $\pm$ 3.9	59.5 $\pm$ 5.9	68.9	59.3 $\pm$ 5.0	56.6 $\pm$ 5.2
ImageNet Supervised	R50	80.3	62.3 $\pm$ 7.4	50.4 $\pm$ 3.8	70.1	57.6 $\pm$ 5.1	50.8 $\pm$ 6.3
ImageNet Supervised	ViT-B	80.4	60.1 $\pm$ 1.4	54.3 $\pm$ 4.4	68.6	62.3 $\pm$ 1.0	58.5 $\pm$ 2.0

Table 2. Stage 1 Results. Bold value represents the highest value for each column. The standard deviation is calculated over the 3 different selections of training data for the low-label settings.

Table 1. We again consider downstream Cholec80 phase recognition and CVS Classification as in Stages 2 and 3.

3 Results and Discussion

Stage 1: Cross-over Experiments. In this stage, we explore several accessible strategies to apply SSL-based pre-training initializations on two public benchmarks (Cholec80 and AutoLaparo) for phase recognition. In Table 2, we see that for the matched pre-trainings, i.e. performing SSL pre-training on the same dataset as the downstream task, there are sizable boosts in performance; particularly so when many labeled examples aren’t available for supervised fine-tuning. Quantitatively, when using as few as 3 labeled videos, we see an 11.3 % and 8.7 % boost over the corresponding ImageNet initialization (R-50) when performing matched pre-training on Cholec80 and AutoLaparo, respectively. In contrast, other than the lowest labeled setting (3 labeled videos), the unmatched pre-trainings perform on par or worse than using corresponding ImageNet initializations (R-50) despite both being laparoscopic video datasets of abdominal procedures. We posit that SSL can help integrate critical representational attributes in low data settings that are otherwise easily learned when more data is available. Further, using the proposed matched finetuning approach on accessible training data (<90k public frames), we outperform EndoFM, the state-of-the-art foundation model for endoscopic data (trained on >5M frames, >70% private) on two public endoscopic benchmarks by upto 2.2% on Cholec80 and 5.1% on AutoLaparo. This highlights the need to carefully consider the application strategy of SSL to both complement and contextualize methodological developments.

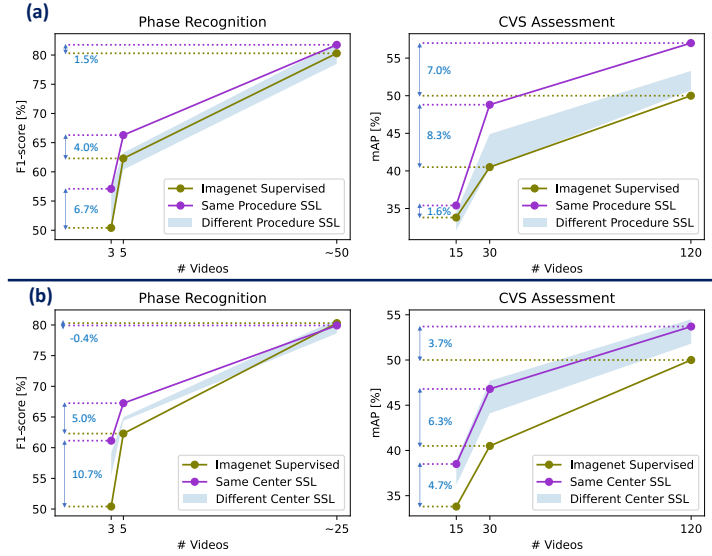


Fig. 2. Phase recognition and CVS assessment results for (a) Stage 2 and (b) Stage 3. In Stage 2, SSL initializations are pre-trained on the relevant (violet line) or various different procedures as the downstream task (shaded blue area). The shaded area represents the range of performance across 7 non-cholecystectomy procedures. In Stage 3, SSL initializations are pre-trained on the relevant (violet line) or various different centers as the downstream task (shaded blue area). The shaded area here represents the range of performance across 6 non-Strasbourg (2 aggregated in HeiChole) centers.

Stage 2: Cross-Procedure Experiments. In Fig. 2, we see that pre-training on the procedure relevant to the downstream tasks (cholecystectomy) brings boosts in performance over ImageNet initializations across the board for all labeled data settings and for both downstream tasks. Interestingly, using the 7 initializations other than those pre-trained on the same procedure as the downstream task delivers performance markedly worse, even lower than ImageNet initializations at times. This key result is both suggestive of the potential of accessible procedure-specific initializations and prompts important questions on how data from other procedures could be leveraged. In Stage 4, we further investigate whether merging subsets of data generated from procedures that are not directly relevant to the downstream task could bring additional value.

Stage 3: Cross-Center Experiments. The stage 3 experiments (Fig. 2 (b)), present a number of interesting findings. First, similar to the previous stage, we see large boosts (up to 10.7 % for phase recognition, 7.2% for CVS assessment) in performance over ImageNet initializations across all data settings and both tasks. Second, when pre-training, domain shifts introduced by sourcing videos from different centers appear to be much less impactful than those introduced by varying procedure types. This effect is resoundingly confirmed across the

Initialization	Cholec80 Phase (F1)			Endoscapes-CVS201 (mAP)		
	Number of Labeled Videos			Number of Labeled Videos		
	40	5	3	120	30	15
StrasChole400 SSL	82.5	67.4 $\pm$ 4.9	62.2 $\pm$ 5.1	52.2	45.2 $\pm$ 4.7	35.0 $\pm$ 7.4
MultiChole2024 SSL	82.2	68.4 $\pm$ 1.3	61.1 $\pm$ 4.5	51.4	46.6 $\pm$ 3.5	34.2 $\pm$ 6.4
Laparo420 (with LC) SSL	79.1	63.9 $\pm$ 3.2	58.1 $\pm$ 6.3	53.3	44.1 $\pm$ 5.4	34.0 $\pm$ 6.6
Laparo420 (w/o LC) SSL	80.5	64.7 $\pm$ 3.3	55.7 $\pm$ 5.7	54.5	41.2 $\pm$ 4.5	31.5 $\pm$ 2.7
ImageNet Supervised	80.3	62.3 $\pm$ 7.4	50.4 $\pm$ 3.8	50.0	40.5 $\pm$ 7.0	33.8 $\pm$ 4.1

Table 3. Stage 4 Results. The standard deviation is calculated over the 3 different selections of training data for the low-label settings.

7 centers (2 aggregated in HeiChole) included in this stage. Unlike the cross-procedure experiments, all initializations, irrespective of which center’s data it was generated using, deliver boosts over the ImageNet initialization, at times even higher than initializations generated using data from the same center. These results suggest that having large-scale procedure-specific initializations such as those generated here, even from different centers, could enable rapid prototyping.

Stage 4: Approaches to Scale. The previous 3 sections present strong evidence that pre-training dataset composition can severely impact downstream performance and must be carefully considered. In this last stage, our goal is to validate findings at scale, understand the interplay between different effects, and offer insights for optimal scaling of SSL in surgical computer vision. In Table 3, when using initializations pre-trained on large (~ 400 videos) cholecystectomy datasets (StrasChole400 and MultiCholec2024), we see performance largely on par with the cholecystectomy-only datasets included in previous study stages despite being an order of magnitude larger. An important caveat to this finding, though consistent with previous works [17], is that higher capacity models may be better suited, if not required, to effectively leverage this scale of data for self-supervised pre-training [6]. Finally, we aim to provide indications about the impact of curating multi-procedure datasets for SSL pre-training using two ~ 400 -video scale datasets, Laparo420 (with LC) and Laparo420 (w/o LC), containing 45/420 and 0/375 cholecystectomy procedures, respectively. Both datasets, despite being of comparable size with StrasChole400 and MultiCholec2024, perform markedly worse, reinforcing the findings of Stage 2 that learning procedure-agnostic initializations may require dedicated methodological design. To the best of our knowledge, each of the 3 previous works utilizing massive scale pre-training [1, 6, 22] have only validated their pre-trained initializations on downstream tasks well-represented in their pre-training data. These findings further emphasize the need for more data-centric studies as we scale SSL methodology toward building a true procedure-agnostic foundation model for surgery.

4 Conclusion

Recent developments in self-supervised learning for surgery allow us to learn generic representational features, useful for a range of downstream tasks in the

domain. In this work, we build on these developments, shifting the focus from models to data, and explore whether pre-trained networks could be an effective means to consolidate disparate datasets in the field of Surgical Data Science. We try to illustrate, that simple model initializations could help promote work on new problems in new procedures at new clinical centers with fewer annotations. The pre-trained initializations generated through this work will be made available alongside the released code at <https://github.com/CAMMA-public/ScalingSurgicalSSL/>.

5 Acknowledgement

This work was supported by French state funds managed by the ANR under Grant ANR-20-CHIA-0029-01 (Chair AI4ORSafety) and Grant ANR-10-IAHU-02 (IHU Strasbourg), and by BPI France under reference DOS0180017/00 (project 5G-OR). The AI4SafeChole consortium is represented by Ludovica Baldari, Giovanni Guglielmo Laracca, Ludovica Guerriero, Segio Alfieri, Claudio Fiorillo, Giuseppe Quero, Elisa Cassinotti, Luigi Boni, Diego Cuccurullo, Guido Costamagna, and Bernard Dallemagne.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Batić, D., Holm, F., Özsoy, E., Czempiel, T., Navab, N.: Endovit: pretraining vision transformers on a large collection of endoscopic images. *International Journal of Computer Assisted Radiology and Surgery* **19**(6), 1085–1091 (2024)
2. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297* (2020)
3. da Costa Rocha, C., Padoy, N., Rosa, B.: Self-supervised surgical tool segmentation using kinematic information. In: *2019 International Conference on Robotics and Automation (ICRA)*. pp. 8720–8726. IEEE (2019)
4. Czempiel, T., Paschali, M., Keicher, M., Simson, W., Feussner, H., Kim, S.T., Navab, N.: Tecno: Surgical phase recognition with multi-stage temporal convolutional networks. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III* 23. pp. 343–352. Springer (2020)
5. Funke, I., Jenke, A., Mees, S.T., Weitz, J., Speidel, S., Bodenstedt, S.: Temporal coherence-based self-supervised learning for laparoscopic workflow analysis. In: *International Workshop on Computer-Assisted and Robotic Endoscopy*. pp. 85–93. Springer (2018)
6. Hirsch, R., Caron, M., Cohen, R., Livne, A., Shapiro, R., Golany, T., Goldenberg, R., Freedman, D., Rivlin, E.: Self-supervised learning for endoscopic video analysis. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 569–578. Springer (2023)
7. Kannan, S., Yengera, G., Mutter, D., Marescaux, J., Padoy, N.: Future-state predicting lstm for early surgery type recognition. *IEEE Transactions on Medical Imaging* **39**(3), 556–566 (2019)

8. Kassem, H., Alapatt, D., Mascagni, P., AI4SafeChole, C., Karargyris, A., Padoy, N.: Federated cycling (fedcy): Semi-supervised federated learning of surgical phases. *IEEE Transactions on Medical Imaging* (2022)
9. Madani, A., Namazi, B., Altieri, M.S., Hashimoto, D.A., Rivera, A.M., Pucher, P.H., Navarrete-Welton, A., Sankaranarayanan, G., Brunt, L.M., Okrainec, A., et al.: Artificial intelligence for intraoperative guidance: using semantic segmentation to identify surgical anatomy during laparoscopic cholecystectomy. *Annals of surgery* **276**(2), 363–369 (2022)
10. Maier-Hein, L., Eisenmann, M., Sarikaya, D., März, K., Collins, T., Malpani, A., Fallert, J., Feussner, H., Giannarou, S., Mascagni, P., et al.: Surgical data science—from concepts toward clinical translation. *Medical image analysis* **76**, 102306 (2022)
11. Maier-Hein, L., Vedula, S.S., Speidel, S., Navab, N., Kikinis, R., Park, A., Eisenmann, M., Feussner, H., Forestier, G., Giannarou, S., et al.: Surgical data science for next-generation interventions. *Nature Biomedical Engineering* **1**(9), 691–696 (2017)
12. Mascagni, P., Alapatt, D., Lapergola, A., Vardazaryan, A., Mazellier, J.P., Dallemagne, B., Mutter, D., Padoy, N.: Early-stage clinical evaluation of real-time artificial intelligence assistance for laparoscopic cholecystectomy. *British Journal of Surgery* **111**(1), znad353 (2024)
13. Mascagni, P., Alapatt, D., Urade, T., Vardazaryan, A., Mutter, D., Marescaux, J., Costamagna, G., Dallemagne, B., Padoy, N.: A computer vision platform to automatically locate critical events in surgical videos: documenting safety in laparoscopic cholecystectomy. *Annals of surgery* **274**(1), e93–e95 (2021)
14. Mascagni, P., Vardazaryan, A., Alapatt, D., Urade, T., Emre, T., Fiorillo, C., Pessaux, P., Mutter, D., Marescaux, J., Costamagna, G., et al.: Artificial intelligence for surgical safety: automatic assessment of the critical view of safety in laparoscopic cholecystectomy using deep learning. *Annals of surgery* **275**(5), 955–961 (2022)
15. Murali, A., Alapatt, D., Mascagni, P., Vardazaryan, A., Garcia, A., Okamoto, N., Costamagna, G., Mutter, D., Marescaux, J., Dallemagne, B., et al.: The endoscapes dataset for surgical scene segmentation, object detection, and critical view of safety assessment: Official splits and benchmark. *arXiv preprint arXiv:2312.12429* (2023)
16. Neimark, D., Bar, O., Zohar, M., Hager, G.D., Asselmann, D.: “train one, classify one, teach one”-cross-surgery transfer learning for surgical step recognition. In: *Medical Imaging with Deep Learning*. pp. 532–544. PMLR (2021)
17. Ramesh, S., Srivastav, V., Alapatt, D., Yu, T., Murali, A., Sestini, L., Nwoye, C.I., Hamoud, I., Sharma, S., Fleurentin, A., et al.: Dissecting self-supervised learning methods for surgical computer vision. *Medical Image Analysis* **88**, 102844 (2023)
18. Ross, T., Zimmerer, D., Vemuri, A., Isensee, F., Wiesenfarth, M., Bodenstedt, S., Both, F., Kessler, P., Wagner, M., Müller, B., et al.: Exploiting the potential of unlabeled endoscopic video data with self-supervised learning. *International journal of computer assisted radiology and surgery* **13**, 925–933 (2018)
19. Sestini, L., Rosa, B., De Momi, E., Ferrigno, G., Padoy, N.: A kinematic bottleneck approach for pose regression of flexible surgical instruments directly from images. *IEEE Robotics and Automation Letters* **6**(2), 2938–2945 (2021)
20. 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 transactions on medical imaging* **36**(1), 86–97 (2016)
21. Wagner, M., Müller-Stich, B.P., Kisilenko, A., Tran, D., Heger, P., Mündermann, L., Lubotsky, D.M., Müller, B., Davitashvili, T., Capek, M., et al.: Comparative

- validation of machine learning algorithms for surgical workflow and skill analysis with the heichole benchmark. *Medical Image Analysis* p. 102770 (2023)
22. Wang, Z., Liu, C., Zhang, S., Dou, Q.: Foundation model for endoscopy video analysis via large-scale self-supervised pre-train. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 101–111. Springer (2023)
 23. Wang, Z., Lu, B., Long, Y., Zhong, F., Cheung, T.H., Dou, Q., Liu, Y.: Autolaparo: A new dataset of integrated multi-tasks for image-guided surgical automation in laparoscopic hysterectomy. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 486–496. Springer (2022)