Imitation Learning from Expert Video Data for Dissection Trajectory Prediction in Endoscopic Surgical Procedure

Jianan Li[1], Yueming Jin[2], Yueyao Chen[1], Hon-Chi Yip[3], Markus Scheppach[4], Philip Wai-Yan Chiu[5], Yeung Yam[6], Helen Mei-Ling Meng[7], and Qi Dou[1](✉)

[1] Department of Computer Science and Engineering,
The Chinese University of Hong Kong, Hong Kong, China
qidou@cuhk.edu.hk

[2] Department of Biomedical Engineering and Department of Electrical and Computer Engineering, National University of Singapore, Singapore, Singapore

[3] Department of Surgery, The Chinese University of Hong Kong, Hong Kong, China

[4] Internal Medicine III - Gastroenterology,
University Hospital of Augsburg, Augsburg, Germany

[5] Multi-scale Medical Robotics Center and The Chinese University of Hong Kong, Hong Kong, China

[6] Department of Mechanical and Automation Engineering,
The Chinese University of Hong Kong, Hong Kong, China

[7] Centre for Perceptual and Interactive Intelligence and The Chinese University of Hong Kong, Hong Kong, China

Abstract. High-level cognitive assistance, such as predicting dissection trajectories in Endoscopic Submucosal Dissection (ESD), can potentially support and facilitate surgical skills training. However, it has rarely been explored in existing studies. Imitation learning has shown its efficacy in learning skills from expert demonstrations, but it faces challenges in predicting uncertain future movements and generalizing to various surgical scenes. In this paper, we introduce imitation learning to the formulated task of learning how to suggest dissection trajectories from expert video demonstrations. We propose a novel method with implicit diffusion policy imitation learning (iDiff-IL) to address this problem. Specifically, our approach models the expert behaviors using a joint state-action distribution in an implicit way. It can capture the inherent stochasticity of future dissection trajectories, therefore allows robust visual representations for various endoscopic views. By leveraging the diffusion model in policy learning, our implicit policy can be trained and sampled efficiently for accurate predictions and good generalizability. To achieve conditional sampling from the implicit policy, we devise a forward-process guided action inference strategy that corrects the state mismatch. We collected a private ESD video dataset with 1032 short clips to validate our method. Experimental results demonstrate that our solution outperforms SOTA imitation learning methods on our formulated task. To the best of our knowledge, this is the first work applying imitation learning for surgical skill learning with respect to dissection trajectory prediction.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14228, pp. 494–504, 2023.
https://doi.org/10.1007/978-3-031-43996-4_47

Keywords: Imitation Learning · Surgical Trajectory Prediction · Endoscopic Submucosal Dissection · Surgical Data Science

1 Introduction

Despite that deep learning models have shown success in surgical data science to improve the quality of surgical intervention [20–22], such as intelligent workflow analysis [7,13] and scene understanding [1,28], research on higher-level cognitive assistance for surgery still remains underexplored. One essential task is supporting decision-making on dissection trajectories [9,24,29], which is challenging yet crucial for ensuring surgical safety. Endoscopic Submucosal Dissection (ESD), a surgical procedure for treating early gastrointestinal cancers [2,30], involves multiple dissection actions that require considerable experience to determine the optimal dissection trajectory. Informative suggestions for dissection trajectories can provide helpful cognitive assistance to endoscopists, for mitigation of intraoperative errors, reducing risks of complications [15], and facilitating surgical skill training [17]. However, predicting the desired trajectory for future time frames based on the current endoscopic view is challenging. First, the decision of dissection trajectories is complicated and depends on numerous factors such as safety margins surrounding the tumor. Second, dynamic scenes and poor visual conditions may further hamper scene recognition [27]. To date, there is still no work on data-driven solutions to predict such dissection trajectories, but we argue that it is possible to reasonably learn this skill from expert demonstrations based on video data.

Imitation learning has been widely studied in various domains [11,16,18] with its good ability to learn complex skills, but it still needs adaptation and improvement when being applied to learn dissection trajectory from surgical data. One challenge arises from the inherent uncertainty of future trajectories. Supervised learning such as Behavior Cloning (BC) [3] tends to average all possible prediction paths, which leads to inaccurate predictions. While advanced probabilistic models are employed to capture the complexity and variability of dissection trajectories [14,19,25], how to ensure reliable predictions across various surgical scenes still remains a great challenge. To overcome these issues, implicit models are emerging for policy learning, inspiring us to rely on implicit Behavior Cloning (iBC) [5], which can learn robust representations by capturing the shared features of both visual inputs and trajectory predictions with a unified implicit function, yielding superior expressivity and visual generalizability. However, these methods still bear their limitations. For instance, approaches leveraging energy-based models (EBMs) [4–6,12] suffer from intensive computations due to reliance on the Langevin dynamics, which leads to a slow training process. In addition, the model performance can be sensitive to data distribution and the noise in training data would result in unstable trajectory predictions.

In this paper, we explore an interesting task of predicting dissection trajectories in ESD surgery via imitation learning on expert video data. We propose Implicit Diffusion Policy Imitation Learning (iDiff-IL), a novel imitation learning

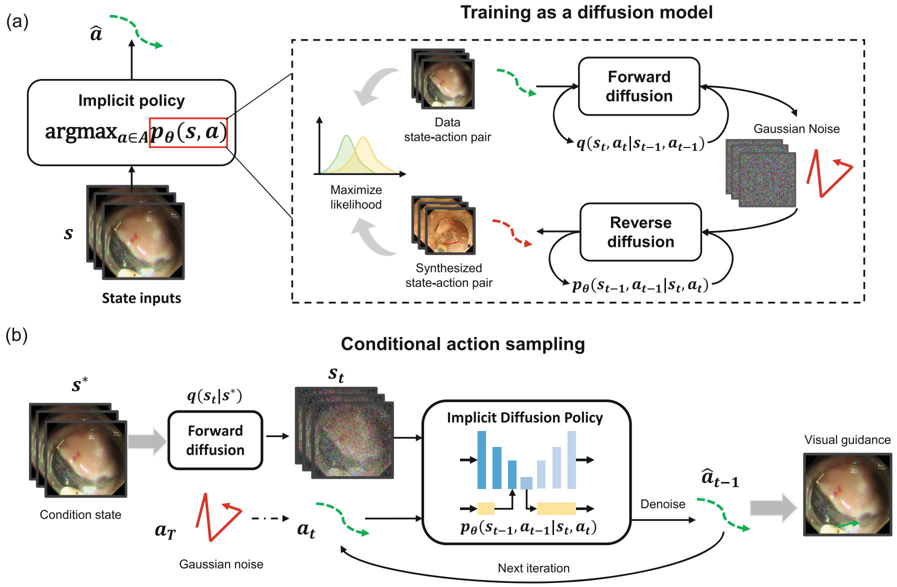


Fig. 1. Overview of our imitation learning method for surgical dissection trajectory prediction. (a) illustrates the modeling of the implicit policy, and its training process. We train a diffusion model to approximate the joint state-action distribution, and (b) depicts the inference loop for trajectory prediction with the forward-diffusion guidance.

approach for dissection trajectory prediction. To effectively model the surgeon's behaviors and handle the large variation of surgical scenes, we leverage implicit modeling to express expert dissection skills. To address the limitations of inefficient training and unstable performance associated with EBM-based implicit policies, we formulate the implicit policy using an unconditional diffusion model, which demonstrates remarkable ability in representing complex high-dimensional data distribution for videos. Subsequently, to obtain predictions from the implicit policy, we devise a conditional action inference strategy with the guidance of forward-diffusion, which further improves the prediction accuracy. For experimental evaluation, we collected a surgical video dataset of ESD procedures, and preprocessed 1032 short clips with dissection trajectories labelled. Results show that our method achieves superior performances in different contexts of surgical scenarios compared with representative popular imitation learning methods.

2 Method

In this section, we describe our approach iDiff-IL, which learns to predict the dissection trajectory from expert video data using the implicit diffusion policy. An overview of our method is shown in Fig. 1. We first present the formulation of the task and the solution with implicit policy for dissection trajectory learning. Next, we present how to train the implicit policy as an unconditional generative diffusion model. Finally, we show the action inference strategy with forward-diffusion guidance which produces accurate trajectory predictions with our implicit diffusion policy.

2.1 Implicit Modeling for Surgical Dissection Decision-Making

In our approach, we formulate the dissection trajectory prediction to an imitation learning from expert demonstrations problem, which defines a Markov Decision Process (MDP) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{D})$, comprising of state space $\mathcal{S}$, action set $\mathcal{A}$, state transition distribution $\mathcal{T}$, and expert demonstrations $\mathcal{D}$. The goal is to learn a prediction policy $\pi^*(a|s)$ from a set of expert demonstrations $\mathcal{D}$. The input state of the policy is a clip of video frames $s = \{I_{t-L+1}, I_{t-L+2}, \ldots, I_t\}$, $I_t \in \mathbb{R}^{H \times W \times 3}$ and the output is an action distribution of a sequence of 2D coordinates $a = \{y_{t+1}, y_{t+2}, ..., y_{t+N}\}, y_t \in \mathbb{R}^2$ indicating the future dissection trajectory projected to the image space.

In order to obtain the demonstrated dissection trajectories from the expert video data, we first manually annotate the dissection trajectories on the video frame according to the moving trend of the instruments observed from future frames, then create a dataset $\mathcal{D} = \{(s, a)_i\}_{i=0}^{M}$ containing M pairs of video clip (state) and dissection trajectory (action).

To precisely predict the expert dissection behaviors and effectively learn generalizable features from the expert demonstrations, we use the implicit model as our imitation policy. Extending the formulation in [5], we model the dissection trajectory prediction policy to a maximization of the joint state-action probability density function $\arg\max_{a \in \mathcal{A}} p_\theta(s, a)$ instead of an explicit mapping $F_\theta(s)$. The optimal action is derived from the policy distribution conditioned on the state s, and $p_\theta(s, a)$ represents the joint state-action distribution.

To learn the implicit policy from the demonstrations, we adopt the Behavior Cloning objective which is to essentially minimize the Kullback-Leibler (KL) divergence between the learning policy $\pi_\theta(a|s)$ and the demonstration distribution $\mathcal{D}$, also equivalent to maximize the expected log-likelihood of the joint state-action distribution, as shown:

$$\max_\theta \mathbb{E}_{(s,a)\sim\mathcal{D}}[\log \pi_\theta(a|s)] = \max_\theta \mathbb{E}_{(s,a)\sim\mathcal{D}}[\log p_\theta(s, a)]. \qquad (1)$$

In this regard, the imitation of surgical dissection decision-making is converted to a distribution approximation problem.

2.2 Training Implicit Policy as Diffusion Models

Approximating the joint state-action distribution in Eq. 1 from the video demonstration data is challenging for previous EBM-based methods. To address the learning of implicit policy, we rely on recent advances in diffusion models. By representing the data using a continuous thermodynamics diffusion process, which can be discretized into a series of Gaussian transitions, the diffusion model is able to express complex high-dimensional distribution with simple parameterized functions. In addition, the diffusion process also serves as a form of data augmentation by adding a range of levels of noise to the data, which guarantees a better generalization in high-dimensional state space.

As shown in Fig. 1 (a), the diffusion model comprises a predefined forward diffusion process and a learnable reverse denoising process. The forward process gradually diffuses the original data $x_0 = (s, a)$, to a series of noised data $\{x_0, x_1, \cdots, x_T\}$ with a Gaussian kernel $q(x_t|x_{t-1})$, where T denotes the diffusion step. In the reverse process, the data is recovered via a parameterized Gaussian $p_\theta(x_{t-1}|x_t)$ iteratively. With the reverse process, the joint state-action distribution in the implicit policy can be expressed as:

$$p_\theta(x_0) = \sum_{x_{1:T}} p_\theta(x_{0:T}) = \sum_{x_{1:T}} p(x_T) \prod p_\theta(x_{t-1}|x_t) = \mathbb{E}_{p_\theta(x_{1:T})} p_\theta(x_0|x_1). \quad (2)$$

The probability of the noised data x_t in forward diffusion process is a Gaussian distribution expressed as $q(x_t|x_0) = \mathcal{N}(x_t, \sqrt{\alpha_t}x_0, (1-\alpha_t)\boldsymbol{I})$, where α_t is a scheduled variance parameter, which can be referred from [10], and $\boldsymbol{I}$ is an identity matrix. The trainable reverse transition is a Gaussian distribution as well, whose posterior is $p_\theta(x_{t-1}|x_t) = \mathcal{N}\big(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)\big)$, in which $\mu_\theta(x_t, t)$ and $\Sigma_\theta(x_t, t)$ are the means and the variances parameterized by a neural network.

To train the implicit diffusion policy, we maximize the log-likelihood of the state-action distribution in Eq. 1. Using the Evidence Lower Bound (ELBO) as the proxy, the likelihood maximization can be simplified to a noise prediction problem, more details can be referred to [10]. Noise prediction errors for the state and the action are combined using a weight $\gamma \in [0, 1]$ as the following:

$$\mathcal{L}_{noise}(\theta) = \mathrm{E}_{\epsilon,t,x_0}[(1-\gamma)\|\epsilon_\theta^a(x_t, t) - \epsilon^a\| + \gamma\|\epsilon_\theta^s(x_t, t) - \epsilon^s\|], \quad (3)$$

where ϵ^s and ϵ^a are sampled from $\mathcal{N}(0, \boldsymbol{I}^s)$, $\mathcal{N}(0, \boldsymbol{I}^a)$ respectively. To better process features from video frames and trajectories of coordinates, we employ a variant of the UNet as the implicit diffusion policy network, where the trajectory information is fused into feature channels via MLP embedding layers. Then the trajectory noise is predicted by an MLP branch at the bottleneck layer.

2.3 Conditional Sampling with Forward-Diffusion Guidance

Since the training process introduced in Sect. 2.2 is for unconditional generation, the conventional sampling strategy through the reverse process will predict random trajectories in expert data. An intuitive way to introduce the condition into the inference is to input the video clip as the condition state s^* to the implicit diffusion policy directly, then only sample the action part. But there is a mismatch between the distribution of the state s^* and the s_t in the training process, which may lead to inaccurate predictions. Hence, we propose a sampling strategy to correct such distribution mismatch by introducing the forward-process guidance into the reverse sampling procedure.

Considering the reverse process of the diffusion model, the transition probability conditioned by s^* can be decomposed as:

$$p_\theta(x_{t-1}|x_t, s^*) = p_\theta(x_{t-1}|s_t, a_t, s^*) = p_\theta(x_{t-1}|x_t)q(s_t|s^*), \tag{4}$$

where $x_t = (s_t, a_t)$, $p_\theta(x_{t-1}|x_t)$ denotes the learned denoising function of the implicit diffusion model, and $q(s_t|s^*)$ represents a forward diffusion process from the condition state to the t-th diffused state. Therefore, we can attain conditional sampling via the incorporation of forward-process guidance into the reverse sampling process of the diffusion model.

The schematic illustration of our sampling approach is shown in Fig. 1 (b). At the initial step $t = T$, action a_T is sampled from a pure Gaussian noise, whereas the input state s_T is diffused from the input video clip s^* through a forward-diffusion process. At the t-th step of the denoising loop, the action input a_t comes from the denoised action from the last time step, while the visual inputs s_t are still obtained from s^* via the forward diffusion process. The above forward diffusion process and the denoising step are repeated till $t = 0$. The final action $\hat{a}_0$ is the prediction from the implicit diffusion policy. The deterministic action can be obtained by taking the most probable samples during the reverse process.

3 Experiments

3.1 Experimental Dataset and Evaluation Metrics

Dataset. We evaluated the proposed approach on a dataset assembled from 22 videos of ESD surgery cases, which are collected from the Endoscopy Centre of the Prince of Wales Hospital in Hong Kong. All videos were recorded via Olympus microscopes operated by an expert surgeon with over 15 years of experience in ESD. Considering the inference speed, we downsampled the original videos to 2FPS frames which are resized to 128×128 in resolution. The input state is a 1.5-s length video clip containing 3 consecutive frames, and the expert dissection trajectory is represented by a 6-point polyline indicating the tool's movements in future 3 s. We totally annotated 1032 video clips, which contain 3 frames for each clip. We randomly selected 742 clips from 20 cases for training, consisting of 2226 frames, where 10% of these are for validation. The remaining 290 clips (consisting of 970 frames) were used for testing.

Experiment Setup. First, to study how the model performs on data within the same surgical context as the training data, we define a subset, referred as to the "in-the-context" testing set, which consists of consecutive frames selected from the same cases as included in the training data. Second, to assess the model's ability to generalize to visually distinct scenes, we created an "out-of-the-context" testing set that is composed of video clips sampled from 2 unseen surgical cases. The sizes of these two subsets are 224 and 66 clips, respectively.

Evaluation Metrics. To evaluate the performance of the proposed approach, we adopt several metrics, including commonly used evaluation metrics for trajectory prediction as used in [23,26], including Average Displacement Error (ADE),

Table 1. Quantitative results on the in-the-context and the out-of-the-context data in metrics of ADE/FDE/FD. Values in parentheses denote video-clip wise standard deviation. The lower is the better.

Method	In-the-context						Out-of-the-context					
	ADE		FDE		FD		ADE		FDE		FD	
BC	10.43	(±5.52)	16.68	(±10.59)	24.92	(±14.59)	13.67	(±6.43)	17.03	(±12.50)	29.23	(±16.56)
iBC[5]	15.54	(±4.79)	22.66	(±8.06)	35.26	(±11.78)	15.81	(±4.66)	19.66	(±7.56)	31.66	(±9.14)
MID[8]	9.90	(±0.66)	15.26	(±1.35)	23.78	(±1.73)	12.42	(±1.89)	16.32	(±3.11)	27.04	(±4.79)
Ours	**9.47**	(±1.66)	**13.85**	(±2.01)	**21.43**	(±3.89)	**10.21**	(±3.17)	**14.14**	(±3.63)	**21.56**	(±5.97)

which respectively reports the overall deviations between the predictions and the ground truths, and Final Displacement Error (FDE) describing the difference from the moving target by computing the L2 distance between the last trajectory points. Besides, we also use the Fréchet Distance (FD) metric, to indicate the geometrical similarity between two temporal sequences. Pixel errors are used as units for all metrics, while the input images are in 128×128 resolution.

3.2 Comparison with State-of-the-Art Methods

To evaluate the proposed approach, we have selected popular baselines and state-of-the-art methods for comparison. We have chosen the fully supervised method, Behavior Cloning, as the baseline, which is implemented using a CNN-MLP network. In addition, we have included iBC [5], an EBM-based implicit policy learning method and MID [8], a diffusion-based trajectory prediction approach, as comparison state-of-the-art approaches.

As shown in Table 1, our method outperforms the comparison approaches in both "in-the-context" and "out-of-the-context" scenarios on all metrics. Compared with the diffusion-based method MID [8], our iDiff-IL is more effective in predicting long-term goals, particularly in the "out-of-the-context" scenes, with the evidence of 2.18 error reduction on FDE. For iBC [5], the performance did not meet our expectations and was even surpassed by the baseline. This exhibits the limitations of EBM-based methods in learning visual representations from complex endoscopic scenes. The superior results achieved by our method demonstrate the effectiveness of the diffusion model in learning the implicit policy from the expert video data. In addition, our method can learn generalizable dissection skills by exhibiting a lower standard deviation of the prediction errors compared to the BC, which severely suffers from over-fitting to the training data. The qualitative results are presented in Fig. 2. We selected three typical scenes in ESD surgery (i.e., submucosa dissection, mucosa dissection and mucosa incision), and showed the predictions of iDiff-IL accompanying the ground truth trajectories. From the results, our method can generate reasonable visual guidance aligning with the expert demonstrations on both evaluation sets.

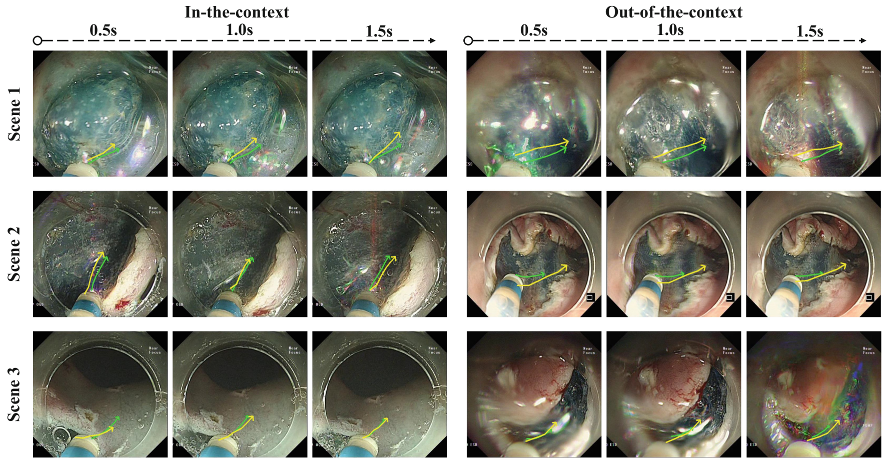


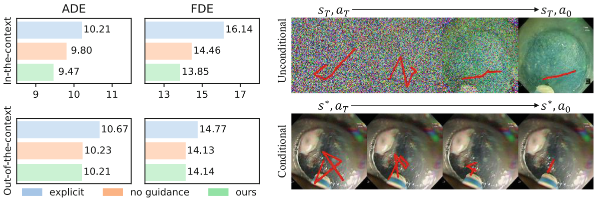
Fig. 2. Typical results of our imitation learning method under settings of in-the-context and out-of-the-context evaluations. Green and yellow arrows respectively denote ground truths and predictions of dissection trajectory. (Color figure online)

3.3 Ablation Study

Implicit Modeling. First, we examined the importance of using implicit modeling as the policy representation. We simulated the explicit form of the imitation policy by training a conditional diffusion model whose conditional input is a video clip. According to the bar charts in Fig. 3, the explicit diffusion policy shows a performance drop for both evaluation sets on ADE compared with the implicit form. The implicit modeling makes a more significant contribution in predicting within the "in-the-context" scenes, suggesting that the implicit model excels at capturing subtle changes in surgical scenes. While our method improves marginally compared with the explicit form on the "out-of-the-context" data, exhibiting a slighter over-fitting with a lower standard deviation.

Forward-Diffusion Guidance. We also investigated the necessity of the forward-diffusion guidance in conditional sampling for prediction accuracy. We remove the forward-diffusion guidance during the action sampling procedure so that the condition state is directly fed into the policy while sampling actions through the reverse process. As shown in Fig. 3, the implicit diffusion policy benefits more from the forward-diffusion guidance in the "in-the-context" scenes, achieving an improvement of 0.33 on ADE. When encountered with the unseen scenarios in "out-of-the-context" data, the performance improvement of such inference strategy is marginal.

Value of Synthetic Data. Since the learned implicit diffusion policy is capable of generating synthetic expert dissection trajectory data, which can potentially reduce the expensive annotation cost. To better explore the value of such synthetic expert data for downstream tasks, we train the baseline model with the generated expert demonstrations. We randomly generated 9K video-trajectory pairs by unconditional sampling from the implicit diffusion policy. Then, we train the BC model with different data, the pure expert data (real), synthetic data only (synt) and the mixed data with the real and the synthetic (mix). The table in Fig. 3 shows the synthetic data is useful as the augmented data for downstream task learning.



Training with synthetic data

Method	In-context	Out-context
BC (real)	10.43 (±5.52)	13.67 (±6.43)
BC (synt)	11.52 (±5.39)	13.51 (±5.37)
BC (mix)	10.05 (±5.31)	12.62 (±5.52)
Ours	**9.47** (±1.66)	**10.21** (±3.17)

Fig. 3. Left: ablation study of key method components; **Middle:** visualization of reverse processes of unconditional/conditional sampling from implicit policy; **Right:** performance of BC trained with synthetic data v.s. our method on ADE.

4 Conclusion

This paper presents a novel approach on imitation learning from expert video data, in order to achieve dissection trajectory prediction in endoscopic surgical procedure. Our iDiff-IL method utilizes a diffusion model to represent the implicit policy, which enhances the expressivity and visual generalizability of the model. Experimental results show that our method outperforms state-of-the-art approaches on the evaluation dataset, demonstrating the effectiveness of our approach for learning dissection skills in various surgical scenarios. We hope that our work can pave the way for introducing the concept of learning from expert demonstrations into surgical skill modelling, and motivate future exploration on higher-level cognitive assistance in computer-assisted intervention.

Acknowledgement. This work was supported in part by Shenzhen Portion of Shenzhen-Hong Kong Science and Technology Innovation Cooperation Zone under HZQB-KCZYB-20200089, in part by Hong Kong Innovation and Technology Commission Project No. ITS/237/21FP, in part by Hong Kong Research Grants Council Project No. T45-401/22-N, in part by Science, Technology and Innovation Commission of Shenzhen Municipality Project No. SGDX20220530111201008.

References

1. Allan, M., et al.: 2018 robotic scene segmentation challenge. arXiv preprint arXiv:2001.11190 (2020)

2. Chiu, P.W.Y., et al.: Endoscopic submucosal dissection (ESD) compared with gastrectomy for treatment of early gastric neoplasia: a retrospective cohort study. Surg. Endosc. **26**, 3584–3591 (2012)
3. Codevilla, F., Santana, E., López, A.M., Gaidon, A.: Exploring the limitations of behavior cloning for autonomous driving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9329–9338 (2019)
4. Du, Y., Mordatch, I.: Implicit generation and modeling with energy based models. In: Advances in Neural Information Processing Systems, vol. 32 (2019)
5. Florence, P., et al.: Implicit behavioral cloning. In: Conference on Robot Learning, pp. 158–168. PMLR (2022)
6. Ganapathi, A., Florence, P., Varley, J., Burns, K., Goldberg, K., Zeng, A.: Implicit kinematic policies: unifying joint and cartesian action spaces in end-to-end robot learning. In: 2022 International Conference on Robotics and Automation (ICRA), pp. 2656–2662. IEEE (2022)
7. Garrow, C.R., et al.: Machine learning for surgical phase recognition: a systematic review. Ann. Surg. **273**(4), 684–693 (2021)
8. Gu, T., et al.: Stochastic trajectory prediction via motion indeterminacy diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 17113–17122 (2022)
9. Guo, J., Sun, Y., Guo, S.: A novel trajectory predicting method of catheter for the vascular interventional surgical robot. In: IEEE International Conference on Mechatronics and Automation, pp. 1304–1309 (2020)
10. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems, vol. 33, pp. 6840–6851 (2020)
11. Hussein, A., Gaber, M.M., Elyan, E., Jayne, C.: Imitation learning: a survey of learning methods. ACM Comput. Surv. **50**(2), 1–35 (2017)
12. Jarrett, D., Bica, I., van der Schaar, M.: Strictly batch imitation learning by energy-based distribution matching. In: Advances in Neural Information Processing Systems, vol. 33, pp. 7354–7365 (2020)
13. Jin, Y., Long, Y., Gao, X., Stoyanov, D., Dou, Q., Heng, P.A.: Trans-svnet: hybrid embedding aggregation transformer for surgical workflow analysis. Int. J. Comput. Assist. Radiol. Surg. **17**(12), 2193–2202 (2022)
14. Ke, L., Choudhury, S., Barnes, M., Sun, W., Lee, G., Srinivasa, S.: Imitation learning as f-divergence minimization. In: Algorithmic Foundations of Robotics XIV: Proceedings of the Fourteenth Workshop on the Algorithmic Foundations of Robotics, vol. 14. pp. 313–329 (2021)
15. Kim, E., et al.: Factors predictive of perforation during endoscopic submucosal dissection for the treatment of colorectal tumors. Endoscopy **43**(07), 573–578 (2011)
16. Kläser, K., et al.: Imitation learning for improved 3D pet/MR attenuation correction. Med. Image Anal. **71**, 102079 (2021)
17. Laurence, J.M., Tran, P.D., Richardson, A.J., Pleass, H.C., Lam, V.W.: Laparoscopic or open cholecystectomy in cirrhosis: a systematic review of outcomes and meta-analysis of randomized trials. HPB **14**(3), 153–161 (2012)
18. Le Mero, L., Yi, D., Dianati, M., Mouzakitis, A.: A survey on imitation learning techniques for end-to-end autonomous vehicles. IEEE Trans. Intell. Transp. Syst. **23**(9), 14128–14147 (2022)
19. Li, Y., Song, J., Ermon, S.: Infogail: interpretable imitation learning from visual demonstrations. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
20. Loftus, T.J., et al.: Artificial intelligence and surgical decision-making. JAMA Surg. **155**(2), 148–158 (2020)

21. Maier-Hein, L., et al.: Surgical data science-from concepts toward clinical translation. Med. Image Anal. **76**, 102306 (2022)
22. Maier-Hein, L., et al.: Surgical data science for next-generation interventions. Nat. Biomed. Eng. **1**(9), 691–696 (2017)
23. Mohamed, A., Qian, K., Elhoseiny, M., Claudel, C.: Social-STGCNN: a social spatio-temporal graph convolutional neural network for human trajectory prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14424–14432 (2020)
24. Qin, Y., Feyzabadi, S., Allan, M., Burdick, J.W., Azizian, M.: Davincinet: joint prediction of motion and surgical state in robot-assisted surgery. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 2921–2928. IEEE (2020)
25. Ren, A., Veer, S., Majumdar, A.: Generalization guarantees for imitation learning. In: Conference on Robot Learning, pp. 1426–1442. PMLR (2021)
26. Sun, J., Jiang, Q., Lu, C.: Recursive social behavior graph for trajectory prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 660–669 (2020)
27. Wang, J., et al.: Real-time landmark detection for precise endoscopic submucosal dissection via shape-aware relation network. Med. Image Anal. **75**, 102291 (2022)
28. Wang, Y., Long, Y., Fan, S.H., Dou, Q.: Neural rendering for stereo 3D reconstruction of deformable tissues in robotic surgery. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 431–441 (2022)
29. Wang, Z., Yan, Z., Xing, Y., Wang, H.: Real-time trajectory prediction of laparoscopic instrument tip based on long short-term memory neural network in laparoscopic surgery training. Int. J. Med. Robot. Comput. Assist. Surg. **18**(6), e2441 (2022)
30. Zhang, J., et al.: Symmetric dilated convolution for surgical gesture recognition. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12263, pp. 409–418. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59716-0_39