See, Predict, Plan: Diffusion for Procedure Planning in Robotic Surgical Videos

Ziyuan Zhao*[1,2], Fen Fang*[1], Xulei Yang†[1], Qianli Xu†[1]
Cuntai Guan[2], and S. Kevin Zhou[3]

[1] Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), Singapore
[2] Nanyang Technological University, Singapore
[3] Suzhou Institute for Advanced Research, University of Science and Technology of China, Suzhou, Jiangsu, 215123, P.R.China

Abstract. Automatic surgical video analysis is pivotal in enhancing the effectiveness and safety of robot-assisted minimally invasive surgery. This study introduces a novel procedure planning task aimed at predicting target-conditioned actions in surgical videos to achieve desired visual goals, thereby addressing the question of "What to do to achieve a desired visual goal?". Leveraging recent advancements in deep learning, particularly diffusion models, our work proposes the Multi-Scale Phase-Condition Diffusion (MS-PCD) framework. This innovative approach incorporates multi-scale visual features into the diffusion process, conditioned by phase class, to generate goal-conditioned plans. By cascading multiple diffusion models with inputs at different scales, MS-PCD adaptively extracts fine-grained visual features, significantly enhancing procedure planning performance in unstructured robotic surgical videos. We establish a new benchmark for procedure planning in robotic surgical videos using the publicly available PSI-AVA dataset, demonstrating that our method notably outperforms existing baselines on several metrics. Our research not only presents an innovative approach to surgical video analysis but also opens new avenues for automation in surgical procedures, contributing to both patient safety and surgical training.

Keywords: Robotic-Assisted Minimally Invasive Surgery · Procedure Planning · Surgical Video Analysis · Diffusion Models

1 Introduction

Automatic surgical video analysis plays a crucial role in robot-assisted minimally invasive surgery (RMIS), which provides surgical context awareness for surgery monitoring [10], intraoperative assistance [17], and decision support [20], promoting operation reliability and patient safety. Recent advancements in deep learning have achieved remarkable success in various surgical video analysis tasks,

* These authors contributed equally to this work.
† These are corresponding authors (yang_xulei@i2r.a-star.edu.sg; qxu@i2r.a-star.edu.sg).

including gesture and workflow recognition [2, 7, 15], and instrument detection and segmentation [14, 32, 33], and phase and frame prediction [8, 9, 29], driving process automation and scene understanding in surgical procedures. However, most of these methods focus on conducting post-hoc analysis, providing an understanding of the current surgical state to answer the question "What is happening?", and recently there are some works proposed to prospectively predict subsequent surgical states, answering "What is going to happen?". Nevertheless, the surgical environment is visually complex and unstructured, leading to uncertainty in future states, and subsequently influencing the goals we expect to achieve. In our work, we take a different view, and introduce a new task as presented in Fig 2, *i.e.,* procedure planning in surgical videos, for predicting target-conditioned actions to answer "What to do to achieve a desired visual goal?". The ability to know what actions to take to achieve the goal is very important for automating surgical robots to plan and conduct complex surgery tasks in clinical scenarios. Moreover, procedure planning in surgical videos can facilitate advanced tasks, such as robot motion control, risk-based alerting, and surgical training [11], benefiting both patients and surgeons.

In recent years, deep learning methods have been applied to address planning from pixel observations for various domains in structured environments, including visual object manipulation [6, 25] and VizDoom navigation [21, 30]. However, it is challenging to learn structured action spaces from unstructured real videos. Recently, Chang *et al.* [3] introduce procedure planning in instructional videos and propose the Dual Dynamics Networks (DDN) model to expedite the latent space learning but suffers from a compounding error. Consequentially, several works based on two-branch autoregressive models have been proposed with different networks, including generative adversarial networks [1] and Transformers [26]. However, such methods involve slow and complex autoregressive processes, requiring visual supervision in intermediate states. Zhao *et al.* [31] propose a single branch nonautoregressive model to simultaneously predict all steps without intermediate visual observations but introduce complex training schemes and complicated inference processes. On the other hand, the diffusion probabilistic model [13, 24] has recently achieved great success in many research fields, such as image synthesis [4, 22], image super-resolution [16], and video generation [12], due to its strong capability in generative modeling. Inspired by this, Wang *et al.* [28] model procedure planning as a conditional sampling process with diffusion model, achieving promising performance in instruction videos without any intermediate supervision.

In contrast to general instruction videos, surgical videos are very visually complex with diverse challenging conditions such as blood, reflection, and motion artifacts, limiting the planning capacity of previous methods in surgical videos. In addition, surgical videos consist of multiple phases and steps with a small inter-class variance and a large intra-class variance, increasing the difficulty of extracting discriminative representations on subtle differences in environments and moving instruments. On the other hand, multiple instruments in use are moved in the narrow field of view, further making the procedure

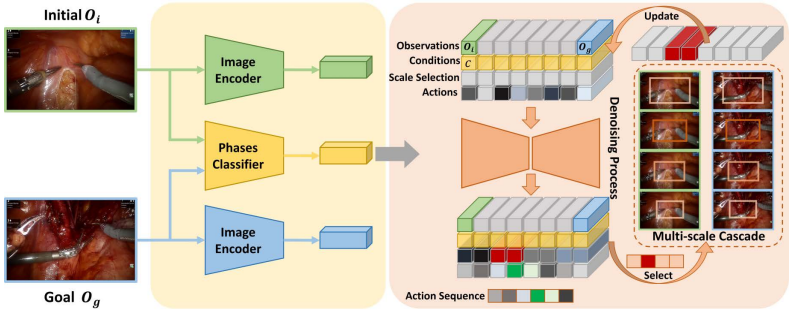

Fig. 1. Overview of our proposed framework. We first generate conditional phase information c by employing a phase classifier, in conjunction with the visual features of initial observations o_i and goal observations o_g for guiding the diffusion process. In the denoising process, we generate multiple scales of inputs and cascade multiple diffusion models that adaptively extract fine-grained visual features for improved performance.

planning task even more difficult in surgical videos. In our work, we propose a novel diffusion-based framework, named Multi-Scale Phase-Condition Diffusion (MS-PCD) for generating goal-conditioned plans given the current visual observations in unstructured robotic surgical videos. Our method incorporates multi-scale visual features into the diffusion process conditioned to phase class by cascading multiple diffusion models with inputs at different scales, thereby extracting more fine-grained visual features adaptively for improved procedure planning performance in surgical videos. To the best of our knowledge, we are the first to investigate the new but challenging procedure planning task in robotic surgical videos. To evaluate the effectiveness of our MS-PCD, we develop a procedure planning benchmark in robotic surgical videos from the publicly available PSI-AVA dataset and compare our method with two strong baselines. Extensive experiments indicate that our method outperforms baselines in surgical videos by a large margin on various metrics for procedure planning.

2 Method

In our surgical procedure planning task, given an initial visual observation o_i and a goal visual state o_g as shown in Fig. 2, our objective is to devise a plan composed of action sequences $\pi = \{a_1, \ldots, a_T\}$, enabling the state transformation from o_i to o_g, where T is the horizon of planning, indicating the number of action steps, and $\{o_i, o_g\}$ indicates two different steps in a surgical event/task, (*i.e.,* phase) of a surgical video. Similar to [28], we decompose surgical procedure planning into two tasks: (1) phase recognition, which is the condition c used to guide step prediction, learning the task-level information based on $\{o_i, o_g\}$, and (2) state-action transition, generating action steps $a_{1:T}$ conditioned on the predicted phase information. In this regard, the surgical procedure planning problem can be expressed as $p(a_{1:T}|o_i, o_g) = \int p(a_{1:T}|o_i, o_g, c)p(c|o_i, o_g)\, dc$. The overview

of the proposed architecture is presented in Fig. 1, which we will discuss in detail in this section.

2.1 Phase-conditioned projected diffusion

Assuming surgical procedure planning as a distribution fitting process solved with a sampling process [28], we propose to leverage the diffusion model for learning the joint distribution of the whole action sequences. Diffusion model [4,13,19] is a class of unconditional generative methods that address the data generation problem by transforming a simple distribution (*e.g.,* Gaussian distribution) into a complex data distribution, consisting of a forward process (or diffusion process) and a reverse process (or denoising process). Given the data distribution $p(y_0)$, the diffusion process is a Markov Chain that incrementally adds Gaussian noise to the input data y_0 for N times:

$$\begin{aligned} q(y_n|y_{n-1}) &= \mathcal{N}(y_n; \sqrt{1-\beta_n}y_{n-1}, \beta_n\mathbf{I}); \\ q(y_n|y_0) &= \mathcal{N}(y_n; \sqrt{\overline{\alpha}_n}y_0, (1-\overline{\alpha}_n)\mathbf{I}), \end{aligned} \tag{1}$$

where β_n is the Gaussian noise ratio at step n, $\overline{\alpha}_n = \prod_{s=1}^{n}(1-\beta_s)$ represents the cumulative effect of noise addition up to step n. The reverse process is modeled as:

$$p(y_{n-1}|y_n) = \mathcal{N}(y_{n-1}; \epsilon_\theta(y_n, t), \sigma_n^2\mathbf{I}), \tag{2}$$

where $\sigma_n^2 = \overline{\beta}_n$, ϵ_θ is trained by a learnable model through an L_2 loss, which can be used to transform noise into data like y_0 after training. Different from the standard diffusion model without guidance for data generation, the distribution fitting process in procedure planning requires the given observations and phase class predicted from the first stage. In this regard, we concatenate (1) visual observations (o_i, o_g), (2) the predicted phase class (c) and (3) candidate action sequence $a_{1:T}$, forming our model input $y = [(o_i, o_g), c, a_{1:T}]$, and employ the condition project scheme with a weight matrix $w = 10$ to emphasize the importance of given visual observations, expressed as:

$$\begin{bmatrix} c_1 & c_2 & & c_T \\ wa_1 & a_2 & \dots & wa_T \\ o_1 & o_2 & & o_T \end{bmatrix} \rightarrow \begin{bmatrix} c & c & & c \\ wa_1 & a_2 & \dots & wa_T \\ o_i & 0 & & o_g \end{bmatrix}, \tag{3}$$

where c_i, a_i and o_i, represent the i^{th} phase class, predicted action logits and observation dimensions in y, respectively. The projection operation $Proj()$ can force the observations and condition dimensions not to change during training and inference. It is well noted that the predicted phase class at the first stage is preliminary but important, especially for visually complex surgical videos. In the regard, we employ the Vision Transformer (ViT) architecture [5] for improving the phase recognition (PR) performance, which outperforms the MLP architecture [28] by 4% in PR accuracy. For the second stage, we follow the basic training

strategy for the diffusion model, but project phases class and observations to the initial values, our training loss is formatted as:

$$\mathcal{L}_{\text{diff}} = \sum_{n=1}^{N} \left(\epsilon_\theta \left(y_n, n\right) - y_0\right)^2. \tag{4}$$

2.2 Multi-scale cascaded diffusion

In the context of surgical videos, diffusion models leverage conditional information and incorporate noise as inputs, showcasing their robustness to noisy environments. However, the small variance among different visual cues and instrument motions within the constrained field of view presents a notable challenge to the diffusion process (Visualizations are shown in Supplementary Material). To overcome these issues, we propose a multi-scale cascaded diffusion process. This approach adaptively selects the optimal input scale from multiple predefined scales (See Supplementary Material for examples of predefined scales), enabling the extraction of more refined visual features, which, in turn, enhances procedure planning performance.

Let S denote the set of predefined scales, with s_i representing the i^{th} scale. During the training phase, the process initiates with a pair of initial and goal observations, $\{o_i, o_g\}$, where the visual embedding, v, is computed as the mean of the visual features extracted from images cropped at predefined scales. For an input image I, the image cropped at scale s_i is represented as I_{s_i}. The visual embedding v for a given epoch is thus defined by:

$$v = \frac{1}{|S|} \sum_{i=1}^{|S|} F(I_{s_i}), \tag{5}$$

where F symbolizes the image encoder. We augment the initial model input by concatenating S with the observation pair and the action sequence $a_{1:T}$, introducing a scale selection branch to the diffusion model. As a result, the action representation in Eq. (3) becomes the concatenation of action sequence and scale selection vector, *i.e.,* replacing a_i with $a_i \oplus s_i$. In the initial 50 steps of each 200-step epoch, the scale s_i is selected randomly from S. Subsequently, the scale selection branch predicts the most suitable scale for cropping. A second diffusion model is then cascaded, processing visual embeddings from images cropped at the predicted optimal scale. Given the lack of concrete ground truth for optimal scale selection, we optimize the scale selection branch indirectly through the overall model objective, avoiding the necessity for additional loss functions and hyperparameters. During inference, the model calculates the visual embedding v using the predefined scales as described. Once trained, it employs D_{ws} (diffusion branch for window scale selection)to determine the optimal cropping window W^*, generating a new visual embedding v' through:

$$v' = F(I_{W^*}). \tag{6}$$

These refined embeddings v' are subsequently fed into the diffusion model for predicting the final sequence of steps. To ensure consistency, the scale selection for both initial and goal observations remains unchanged, reflecting the minimal variation in the active region between these states. Through iterative refinement, involving repeated application of the multi-scale selection and diffusion model, our method adaptively selects the most appropriate scale for the task, thereby improving the accuracy of step sequence predictions.

3 Experiments and Results

Dataset and evaluation metrics. We use the Phase, Step, Instrument, and Atomic Visual Action recognition (PSI-AVA) dataset [27], a public dataset operated with Da Vinci SI3000 Surgical system, consisting of around 20.45 hours of the surgical procedure and $73,618$ keyframes with annotations at the phase and step level, in which $2,238$ keyframes are selected for instrument and atomic action annotations by sampling every 35 seconds. Regarding the experimental framework of procedure planning, we extract all step sequences $\{a_i, \ldots, a_{i+1-T}\}$ with a horizon of planning T from the PSI-AVA dataset, forming sub-tasks with T steps using a sliding window approach [28] (More visual explanations are shown in Supplementary Material).. We set $T = 4$ in our experiments. For each sub-task, the first step and the last step are selected as the initial observation o_i and goal visual state o_g, respectively. We use phase categories as task-level information to guide procedure planning, and the steps in each sub-task belong to the same phase. Finally, we include 10 phases and 20 steps in our evaluation.

To evaluate various methods, we employ three commonly used metrics for procedure planning, *i.e.*, Success Rate (SR), mean Accuracy (mAcc), and mean Intersection over Union (mIoU) [28]. SR considers a plan successful only if all predicted action steps match the ground truth sequence. mACC assesses the correctness of predicted actions at each time step by comparing them with the actions in the ground truth at the same time step. mIoU quantifies the overlap between predicted action steps and ground truth by computing IoU on each individual sequence. In addition, we report the accuracy to evaluate the results for step recognition, Acc-Step for short. All metrics are presented with the mean performance and the variations over multiple inferences, represented as mean($\pm$std).

Baselines. To evaluate the effectiveness of our model in learning plans from surgical videos, we establish strong baselines by employing two state-of-the-art (SOTA) methods designed for procedure planning in instructional Videos *i.e.*, PlaTe [26] and PDPP [28]. PlaTe employs Transformer modules with intermediate visual states, while PDPP uses the diffusion model to model the distribution of sequences of intermediate actions without visual cues. Moreover, we include the TAPIR model [27] for performance comparison on step recognition.

Implementation details. Following [19, 28], we use the cosine noise schedule and the basic U-Net [23] as the learnable model for diffusion. For training, we optimize our model with Adam using a learning rate of $3e-4$ over 100 epochs. The number of training steps per epoch is $N = 200$, and the number of diffusion

Table 1. Performance comparison between our method and state-of-the-art approaches. All results represent the average values obtained from two folds of the PSI-AVA dataset. Values in bracket indicate std of five-fold validation. S is short for scale; $S = n$ means that our scale selection mechanism selects from n predefined scales.

Method	SR (%)↑	mAcc(%)↑	mIOU (%)↑	Acc-Step (%)↑
PlaTe [26]	14.17(±0.07)	23.65(±0.11)	59.32(±0.35)	30.08(±0.18)
PDPP [28]	24.88(±0.14)	36.59(±0.21)	59.44(±0.41)	43.28(±0.35)
TAPIR [27]	-	-	-	39.43(±0.04)
Ours (S=1)	25.04(±0.15)	36.82(±0.25)	59.60(±0.36)	41.38(±0.35)
Ours (S=2)	25.44(±0.18)	37.02(±0.25)	59.66(±0.40)	41.15(±0.36)
Ours (S=3)	25.90(±0.15)	37.24(±0.23)	59.78(±0.38)	42.53(±0.40)
Ours (S=4)	26.76(±0.17)	**39.86(±0.23)**	64.58(±0.40)	**46.47(±0.39)**
Ours (S=5)	**27.05(±0.19)**	39.74(±0.26)	**64.81(±0.43)**	46.20(±0.37)

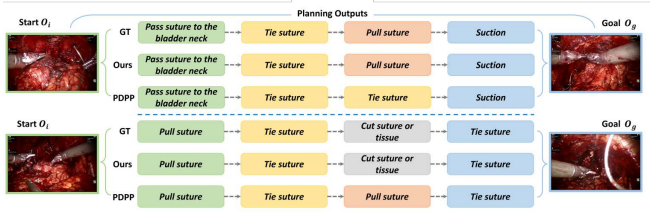

Fig. 2. Visualization of surgical plans produced by our model and PDPP. GT indicates 'Ground Truth'.

steps per training step is 200. We extract action features of 512-dimensional using the encoder trained on the HowTo100M [18] dataset as the input of our diffusion model.

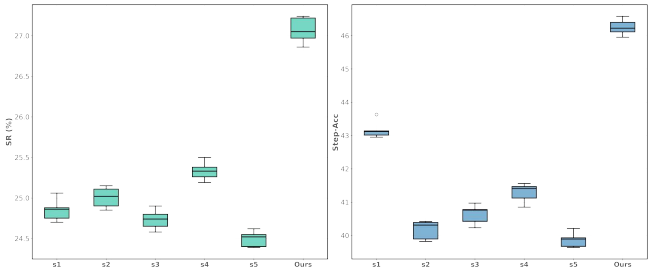

Fig. 3. Boxplot of ablation results with different scales. 's1' to 's5' refer to predefined image scales.

Comparison with other approaches. We present the results of different methods in Table 1. From the table, we can observe that PDPP surpasses Transformer-based PlaTe by a large margin in most metrics, especially for the success rate, which demonstrates the effectiveness of the diffusion model in procedure planning in surgical videos. Notably, our model with different scales achieves better procedure planning performance than other methods by a significant mar-

gin, and outperforms PDPP by 2.17% in SR, 3.27% in mAcc, and 5.37% on mIoU, showing the effectiveness of the proposed method for the task. It is noted that our performance increases dramatically with scales up to 4 in our multi-scale cascaded diffusion process. More specifically, in the procedure planning task (assessed by the first four metrics), adding the fourth scale option contributes to an increase of 0.86% in SR, a 2.62% improvement in mAcc, and a 4.8% rise in mIOU. However, the involvement of the fifth scale shows only marginal gains in some metrics, leading to a plateau in performance improvement. This indicates that while multiple scales can enhance the model performance in capturing fine-grained features for the task, the burden on the scale selection branch would be increased, and affect the selection process, potentially leading to a performance drop on procedure planning. Consequently, we set $S = 5$ in our experiments.

For step recognition, we can see that our method (S=4) outperforms TAPIR by 7% in Acc-Step, suggesting that our method can significantly enhance step recognition performance when aligning the input formation with TAPIR. Furthermore, we show some qualitative results of PDPP and our method in Fig. 2 with more visual comparisons shown in Supplementary Material. From the figure, we can perceive that our model can generate reasonable and diverse plans with the same initial and goal observations, while it is difficult for PDPP to capture the subtle changes in visual cues of surgical videos, leading to failure in procedure planning.

Ablation study. To further demonstrate the efficacy of our proposed multi-scale cascaded diffusion process, we conduct experiments utilizing each predefined scale individually within the diffusion model, as depicted in Fig. 3. The results reveal that employing the 4th scale can yield the best performance, indicating that this scale possesses richer fine-grained features in most of samples to the procedure planning task. When involving the multi-scale selection with our proposed method, better performance can be obtained, which demonstrates the effectiveness of the proposed method in enhancing action sequence prediction in surgical videos.

4 Conclusions

Our investigation into procedure planning in robotic surgical videos introduces the innovative Multi-Scale Phase-Condition Diffusion (MS-PCD) framework, marking a significant advancement in the field of automatic surgical video analysis. By innovatively applying diffusion models conditioned on multi-scale visual features and phase classes, our approach adeptly navigates the complexities of the surgical environment, demonstrating superior performance over existing methodologies. The success of MS-PCD in generating precise, goal-conditioned plans underscores the potential of deep learning in transforming surgical planning and execution. Moreover, the establishment of a new benchmark for evaluating procedure planning underscores the critical need for continued exploration and innovation in this domain.This endeavor not only promises to improve surgical outcomes and patient safety but also to enrich the training and decision-making capabilities of surgeons, heralding a new era of precision and reliability in robot-assisted surgeries.

Acknowledgement. This research work is supported by the Agency for Science, Technology and Research (A*STAR) under its MTC Programmatic Funds. (Grant No. M23L7b0021)

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bi, J., Luo, J., Xu, C.: Procedure planning in instructional videos via contextual modeling and model-based policy learning. In: IEEE CVPR. pp. 15611–15620 (2021)
2. Cao, J., Yip, H.C., Chen, Y., Scheppach, M., Luo, X., Yang, H., Cheng, M.K., Long, Y., Jin, Y., Chiu, P.W.Y., et al.: Intelligent surgical workflow recognition for endoscopic submucosal dissection with real-time animal study. Nature Communications **14**(1), 6676 (2023)
3. Chang, C.Y., Huang, D.A., Xu, D., Adeli, E., Fei-Fei, L., Niebles, J.C.: Procedure planning in instructional videos. In: European Conference on Computer Vision. pp. 334–350. Springer (2020)
4. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
6. Ehsani, K., Han, W., Herrasti, A., VanderBilt, E., Weihs, L., Kolve, E., Kembhavi, A., Mottaghi, R.: Manipulathor: A framework for visual object manipulation. In: IEEE CVPR. pp. 4497–4506 (2021)
7. Funke, I., Bodenstedt, S., Oehme, F., von Bechtolsheim, F., Weitz, J., Speidel, S.: Using 3d convolutional neural networks to learn spatiotemporal features for automatic surgical gesture recognition in video. In: MICCAI. pp. 467–475. Springer (2019)
8. Gao, X., Jin, Y., Dou, Q., Fu, C.W., Heng, P.A.: Accurate grid keypoint learning for efficient video prediction. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 5908–5915. IEEE (2021)
9. Gao, X., Jin, Y., Zhao, Z., Dou, Q., Heng, P.A.: Future frame prediction for robot-assisted surgery. In: International Conference on Information Processing in Medical Imaging. pp. 533–544. Springer (2021)
10. Gubert, L.C., da Costa, C.A., Righi, R.d.R.: Context awareness in healthcare: a systematic literature review. Universal Access in the Information Society **19**, 245–259 (2020)
11. Hashimoto, D.A., Rosman, G., Meireles, O.R.: Artificial intelligence in surgery: understanding the role of AI in surgical practice. McGraw Hill Professional (2021)
12. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., et al.: Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303 (2022)
13. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
14. Jin, Y., Cheng, K., Dou, Q., Heng, P.A.: Incorporating temporal prior from motion flow for instrument segmentation in minimally invasive surgery video. In: MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part V 22. pp. 440–448. Springer (2019)
15. Jin, Y., Long, Y., Chen, C., Zhao, Z., Dou, Q., Heng, P.A.: Temporal memory relation network for workflow recognition from surgical video. IEEE Transactions on Medical Imaging **40**(7), 1911–1923 (2021)
16. Li, H., Yang, Y., Chang, M., Chen, S., Feng, H., Xu, Z., Li, Q., Chen, Y.: Srdiff: Single image super-resolution with diffusion probabilistic models. Neurocomputing **479**, 47–59 (2022)

17. 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)
18. Miech, A., Zhukov, D., Alayrac, J.B., Tapaswi, M., Laptev, I., Sivic, J.: Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In: IEEE CVPR. pp. 2630–2640 (2019)
19. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: International Conference on Machine Learning. pp. 8162–8171. PMLR (2021)
20. Padoy, N.: Machine and deep learning for workflow recognition during surgery. Minimally Invasive Therapy & Allied Technologies **28**(2), 82–90 (2019)
21. Pathak, D., Agrawal, P., Efros, A.A., Darrell, T.: Curiosity-driven exploration by self-supervised prediction. In: International conference on machine learning. pp. 2778–2787. PMLR (2017)
22. Peng, W., Adeli, E., Bosschieter, T., Park, S.H., Zhao, Q., Pohl, K.M.: Generating realistic brain mris via a conditional diffusion probabilistic model. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 14–24. Springer (2023)
23. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015)
24. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International conference on machine learning. pp. 2256–2265. PMLR (2015)
25. Srinivas, A., Jabri, A., Abbeel, P., Levine, S., Finn, C.: Universal planning networks: Learning generalizable representations for visuomotor control. In: International Conference on Machine Learning. pp. 4732–4741. PMLR (2018)
26. Sun, J., Huang, D.A., Lu, B., Liu, Y.H., Zhou, B., Garg, A.: Plate: Visually-grounded planning with transformers in procedural tasks. IEEE Robotics and Automation Letters **7**(2), 4924–4930 (2022)
27. Valderrama, N., Ruiz Puentes, P., Hernández, I., Ayobi, N., Verlyck, M., Santander, J., Caicedo, J., Fernández, N., Arbeláez, P.: Towards holistic surgical scene understanding. In: MICCAI. pp. 442–452. Springer (2022)
28. Wang, H., Wu, Y., Guo, S., Wang, L.: Pdpp: Projected diffusion for procedure planning in instructional videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14836–14845 (2023)
29. Wang, P.: The applications of generative adversarial network in surgical videos. In: Third International Conference on Intelligent Computing and Human-Computer Interaction (ICHCI 2022). vol. 12509, pp. 300–305. SPIE (2023)
30. Wydmuch, M., Kempka, M., Jaśkowski, W.: Vizdoom competitions: Playing doom from pixels. IEEE Transactions on Games **11**(3), 248–259 (2018)
31. Zhao, H., Hadji, I., Dvornik, N., Derpanis, K.G., Wildes, R.P., Jepson, A.D.: P3iv: Probabilistic procedure planning from instructional videos with weak supervision. In: IEEE CVPR. pp. 2938–2948 (2022)
32. Zhao, Z., Jin, Y., Chen, J., Lu, B., Ng, C.F., Liu, Y.H., Dou, Q., Heng, P.A.: Anchor-guided online meta adaptation for fast one-shot instrument segmentation from robotic surgical videos. Medical Image Analysis **74**, 102240 (2021)
33. Zhou, Z., Alabi, O., Wei, M., Vercauteren, T., Shi, M.: Text promptable surgical instrument segmentation with vision-language models. Advances in Neural Information Processing Systems **36** (2024)