2026 IEEE/SICE International Symposium on System Integration (SII)
January 11-14, 2026. Cancún, México

Motion Generation for Surgical Robots using Task-Aware Attention based on Deep Predictive Learning*

Hidetoshi Taira[1], Maina Sogabe[2], Tetsuro Miyazaki[1] and Kenji Kawashima[1]

Abstract— **In this study, we propose a method for autonomously operating a surgical robot by controlling visual attention using the robot's own state in deep predictive learning. The proposed method, named TAIRNN (Task-Attentive Informed Recurrent Neural Network) uses a state-conditioned Query to retrieve visual Keys to deep predictive model. Experimental results of point-to-point movement showed that this method can reach the final target with improved accuracy and in fewer steps compared to conventional methods that rely solely on image information. The results demonstrate that an approach that incorporates a robot's self-state awareness into its visual attention mechanism is effective in suppressing task-irrelevant visual noise and improving control stability.**

I. INTRODUCTION

In recent years, the use of surgical robots has significantly contributed to the precision and minimally invasiveness of surgical procedures, but their operation still relies on the surgeon's hands [1], [2]. In order to reduce the burden on surgeons and standardize advanced surgeries, it is essential to realize autonomous surgery, in which robots perform some of the procedures autonomously [3].

Although it remains difficult to fully automate the entire surgical process, automating subtasks such as tumor resection [4] and suturing [5], [6] is a realistic target task for automation [7], [8]. Deep learning, especially imitation learning that mimics human operations, is a promising method for achieving such autonomous subtasks. However, training surgical robots presents unique challenges. First, collecting high-quality expert data requires the time of skilled surgeons and is costly. Second, the internal environment under endoscopic examination is full of unpredictable factors, such as poor visibility due to bleeding and soft tissue deformation, so the trained model must be extremely robust. Therefore, acquiring a control strategy that can operate stably in such an uncertain environment from a limited dataset is a challenge for realizing autonomous surgery [9], [10], [11], [12]. To address this problem of robust learning with limited data, models based on deep predictive learning frameworks have proven effective [13]. The framework is a self-supervised learning framework in which the model predicts time-series changes in sensory and motor information, including body-environment interactions.

*This work was supported by KAKEN 25H00717.

[1]Hidetoshi Taira, Tetsuro Miyazaki, and Kenji Kawashima are with the Department of Information Physics and Computing, Graduate School of Information Science and Technology, The University of Tokyo, Japan kkawa729@g.ecc.u-tokyo.ac.jp

[2]Maina Sogabe is with the Department of Precision Engineering, Graduate School of Engineering, The University of Tokyo

One of the previous studies that implemented this framework, SARNN (Spatial Attention Recurrent Neural Network), achieved some success by predicting task-important attention points during the reconstruction process of the next image using visual attention [14]. However, the attention mechanism of SARNN relies on salient features in the image and cannot take into account the context of the task, which raises concerns that the model may lack robustness when directly applied to surgical environments [15].

Research has been reported that uses the robot's own state as a cue for attention as a promising approach to overcoming the limitations of attention mechanisms that rely solely on image information. Seneviratne et al. have succeeded in generating adaptive gaits on uneven terrain by combining the visual information of a four-legged robot with inertial and joint information using a cross-attention mechanism [16]. In the field of surgical robotics, Zhao et al., integrate camera images from a bronchoscope with the robot's self-location information using cross-attention, improving the success rate of autonomous operation through reinforcement learning [17]. These studies suggest that integrating physical and visual information of a robot can lead to robust control strategies in uncertain environments.

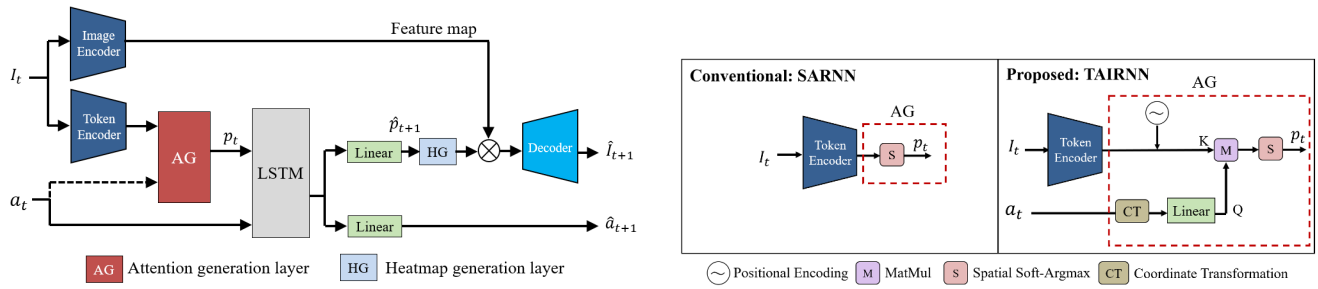
In this study, we propose a time series prediction model called TAIRNN (Task-Attentive Informed Recurrent Neural Network) that incorporates a cross-attention mechanism into the framework of deep predictive learning, which dynamically controls the target of attention using the robot's own state. This allows the model to understand the role of each object in the task and direct its attention towards classification and tracking, rather than relying solely on visual saliency. We demonstrate through a real approach task using a surgical robot that the proposed method improves attentional stability and control accuracy compared to conventional methods.

II. PROPOSED METHOD

This section introduces TAIRNN, a time-series predictive model that extends the Deep Predictive Learning (DPL) framework with a cross-attention mechanism conditioned on the robot's state.

A. System Overview

The architecture of the proposed model is shown in Fig. 1 (a). This model builds upon the framework of the SARNN（Spatial Attention RNN）architecture proposed by



(a) System Overview

(b) Comparison of Attention Mechanisms

Fig. 1. (a) The overall architecture of the model. The Attention Generation (AG) layer extracts attention points p_t from the image I_t. The dashed arrow indicates that, in our proposed method, this process is additionally conditioned on the robot state a_t, which distinguishes it from the conventional approach. (b) Comparison of attention mechanisms. Our proposed method (right) uses a state-conditioned Query (Q) to retrieve visual Keys (K), unlike the conventional method (left), which relies only on image features.

Ichiwara et al. [14], consisting of three main components: Encoder, Recurrent, and Decoder.

- Encoder: The Encoder module is composed of two parallel Convolutional Neural Networks (CNNs): an Image Encoder and a Token Encoder. The Image Encoder extracts general visual feature maps from the camera image I_t at time t for predicting the next frame image. The Token Encoder extracts features for the Attention Generation (AG) mechanism.
- Recurrent Module: The recurrent component employs Long Short-Term Memory (LSTM) architecture. The LSTM module integrates two distinct modalities over time: the attention coordinates p_t (extracted via AG) and the robot state a_t. This integration allows the model to make future predictions based on the spatio-temporal context.
- Decoder: The shared LSTM's hidden state is fed into two separate decoders to predict the next robot state and image frame simultaneously.

    1) Action Decoder: A simple linear layer takes the LSTM's hidden state as input to predict the next robot state $\hat{a}_{t+1}$.
    2) Image Decoder: This decoder first uses a linear layer to predict the attention coordinates for the next frame $\hat{p}_{t+1}$, from the LSTM's hidden state. These coordinates are converted into spatial heatmaps by a Heatmap Generator (HG) layer, which produces localized Gaussian-like responses around each attention point. The resulting heatmaps are multiplied with the feature maps extracted by the Image Encoder to emphasize task-relevant regions. The weighted feature map is then passed through a transposed convolutional network to reconstruct the next image frame, $\hat{I}_{t+1}$. This image prediction process encourages the attention coordinates p_t to learn features relevant to the task, thereby contributing to their stabilization.

B. Attention Mechanism by Cross-Attention

The core innovation of this research lies in the design of the Attention Generation (AG) mechanism. Fig. 1 (b) illustrates a comparison between conventional methods that directly extract attention points from images and our proposed approach using Cross-Attention.

In the conventional method, attention scores are computed solely from an image feature. While this proves effective for extracting visually salient regions within images, this approach has a fundamental limitation: it cannot consider the task context or the robot's current state.

To address this challenge, we introduce a Cross-Attention mechanism. Our proposed method takes as input an image $I_t \in \mathbb{R}^{H\times W\times 3}$ at time step t and a state vector $a_t \in \mathbb{R}^{D_{\text{state}}}$ expressed in the robot arm's coordinate frame. First, a Token Encoder extracts a feature map $F_t \in \mathbb{R}^{H_p\times W_p\times C}$ from the image. Here, H_p, W_p represent the height and width of the feature map, respectively, while C denotes the dimensionality of the feature vector associated with each pixel. This feature map is treated as a set of tokens, where each token corresponds to a spatial location and has a C-dimensional feature vector. To provide the model with absolute spatial information, we add a learnable Positional Encoding to the tokens. Furthermore, Layer Normalization (LN) is applied to stabilize training, producing the set of key vectors $K \in \mathbb{R}^{N\times C}$.

$$K = \text{LN}(\text{Encoder}(I_t) + \text{PositionalEncoding}) \tag{1}$$

Here, $N = H_p \times W_p$ is the total number of tokens.

Concurrently, the Query (Q) vectors are generated from the robot state a_t. First, the end-effector's coordinates are extracted from a_t and transformed from the robot arm's coordinate system to the camera's coordinate system, yielding a'_t through coordinate transformation (CT) using camera pose information. This coordinate alignment ensures consistency between the query and the image feature map (Keys). The resulting vector a'_t is then fed into k independent linear layers to produce k query vectors.

$$Q = \text{Linear}(a'_t) \tag{2}$$

Here, k represents the number of attention points to be extracted and is a task-dependent hyperparameter. In general, an appropriate choice of k depends on factors such as the number and size of task-relevant object, the image resolution, and the characteristics of the task. In this study, we fix $k = 6$ to provide sufficient capacity to capture the relatively large spherical target while keeping the architecture identical across all methods for a fair comparison. A more systematic exploration of k under different tasks and scene conditions is left for future work.

We then compute attention scores by matching the generated Query (Q) against the Key (K) vectors, which are derived from the feature map of the Token Encoder. These scores are passed through a Softmax function to obtain k attention maps $A_t \in \mathbb{R}^{k \times N}$.

$$A_t = \text{softmax}\left(\frac{QK^T}{\sqrt{C}}\right) \tag{3}$$

In this study, we do not explicitly define a value vector V. Instead, we directly use the attention map A_t as the spatial heatmap. The obtained k attention maps A_t are then input into the Spatial Soft-Argmax layer. Similar to conventional approaches, the centroid of each map is extracted, yielding a set of k 2D coordinate vectors $p_t \in \mathbb{R}^{k \times 2}$.

This mechanism enables the model to dynamically determine where and to what extent it should focus attention within the image, based on both the visual input and the robot's current state.

C. Loss Function

Model training is performed through a weighted linear combination of three loss terms. The overall loss function L is defined as follows:

$$L = \lambda_{\text{img}} L_{\text{img}} + \lambda_{\text{state}} L_{\text{state}} + \lambda_{\text{attn}} L_{\text{attn}} \tag{4}$$

Here, L_{img} represents the mean-squared error (MSE) between the predicted and ground-truth next-frame images, L_{state} represents the MSE between the predicted and ground-truth next-frame robot's states, and L_{attn} represents the error of the Euclidean distance between the predicted attention points $\hat{p}_{t+1}$ and target 2D attention points p_t of the Token Encoder. The λ terms are weighting factors that balance the relative importance of each loss term.

III. Experiments

A. Experimental Setup

This experiment quantitatively evaluates the proposed model (TAIRNN) against the conventional model (SARNN) in terms of accuracy and efficiency on a forceps approach task. The experimental setup, shown in Fig. 2, utilizes a prototype of the surgical robot Saroa (Riverfield Inc., Tokyo, Japan) [18]. The robot's end-effector is equipped with a gripper designed to manipulate the target object: a silicone sphere with a diameter of 20 mm, selected for its high visual contrast against the red mat on the workspace. Visual feedback is provided by a fixed-mount endoscope, positioned to offer a view of the forceps tip and the immediate workspace.

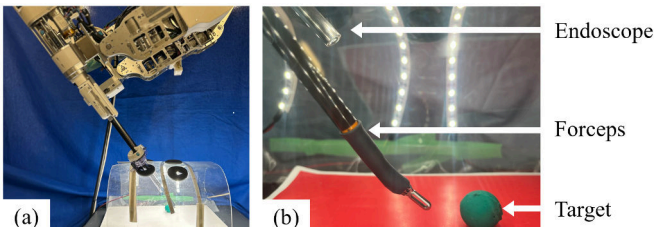


Fig. 2. Experimental setup for the approach task. (a) shows the overall arrangement of the robotic manipulator and workspace. (b) shows a close-up view of the surgical instruments, highlighting the spatial relationship between the forceps (tool) and the silicone sphere (target).

B. Dataset

The training dataset was collected by operating the robot through a complete pick-and-place task, which consists of four phases as shown in Fig. 3. We collected 15 trials, each lasting approximately 8 seconds. At each timestep, we recorded a 64×64 RGB image sampled at approximately 18 Hz and a state vector containing the 3D position and 4D quaternion orientation of the end-effector.

C. Task Definition for Evaluation

Although the training data covers the full pick-and-place sequence, our evaluation focuses solely on the initial "approach" phase, as a stable grasping policy has not yet been implemented. During evaluation, the model performs one inference per control step, operating at 0.35 Hz for a maximum of 100 steps. This rate is set to match the robot's physical execution time; model inference time is not the bottleneck.

To ensure task diversity, we prepared five distinct starting areas on the workspace: a central region and four surrounding corner regions. These areas are represented by the five initial patterns (I-V) shown in Fig. 4.

D. Compared Methods

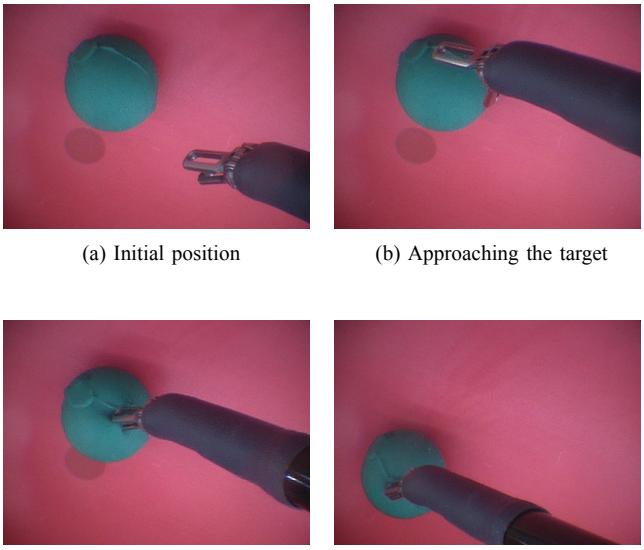
This study compares the following two models:

1. Conventional Method (SARNN): The baseline model employed in previous research [14] that extracts attention points solely from image information.
2. Proposed Method (TAIRNN): The proposed model incorporating a Cross-Attention mechanism that uses queries generated from the robot state a_t.

For a fair comparison, both models were trained on the identical dataset for 10,000 epochs. Key hyperparameters such as learning rate and loss weights were set to values that yielded the most stable performance for each model.

E. Evaluation Metrics

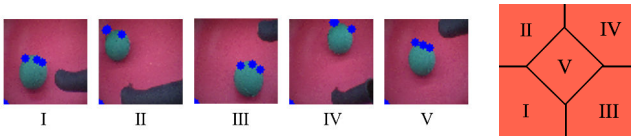
To comprehensively evaluate model performance, we employed the following metrics:



(a) Initial position (b) Approaching the target

(c) Grasping the target (d) Transporting the target

Fig. 3. The four sequential phases of the complete task used for training the proposed model. The experiments in this paper focus specifically on evaluating the performance of the approach phase (b).



(a) Initial Position (b) Initial Areas

Fig. 4. (a) Initial target positions were defined by partitioning the endoscopic view into five regions (I–V); for each region, the score represents the mean performance over three trials. (b) Schematic diagram of the five initial areas for target placement.

- Minimum Distance to Target [mm]: The closest 3D Euclidean distance achieved between the forceps tip and the target's center within a 100-step trial. This metric measures the accuracy of the trajectory.
- Number of Steps to Minimum Distance [steps]: The number of steps taken to reach this minimum distance. This metric measures the efficiency of the model.

IV. Results

A. Overview of Control Performance

Figs. 5 and 6 illustrate the qualitative performance of representative trials for both models. Fig. 5 compares the 3D trajectories of the forceps tip from the same initial position. The conventional method (blue line) attempts a direct, linear approach but stalls before reaching the target area. In contrast, the proposed method (orange line) generates a curved, nonlinear trajectory, successfully getting closer to the target.

This difference is also evident in the distance-over-time plots in Fig. 6, where the proposed model (b) enters the 10 mm target radius while the conventional model (a) does not.

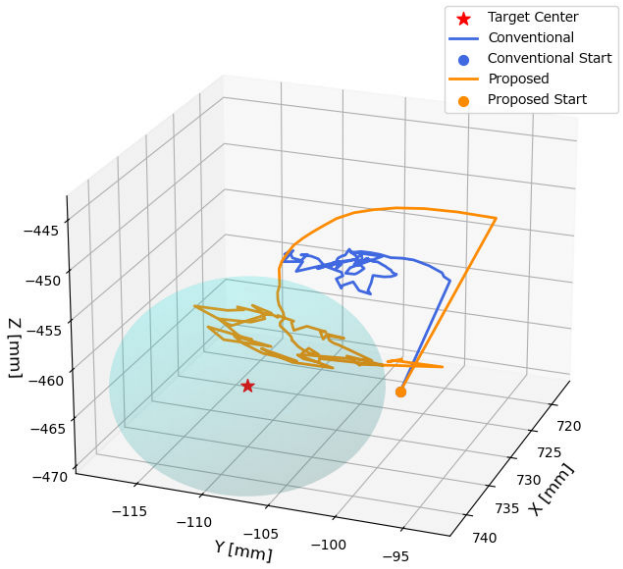


Fig. 5. Representative 3D trajectories of the tool tip for the conventional model (blue) and the proposed model (orange) from the same initial position. The light blue sphere indicates the target area.

B. Analysis of Accuracy and Efficiency

To quantitatively evaluate these performance differences, we analyzed results from all 15 trials (5 patterns × 3 repetitions). Table I summarizes the average Minimum Distance to Target and the corresponding average Steps to Minimum Distance required to reach that distance for both models.

Overall, the proposed method demonstrates superior performance in both accuracy and efficiency. For accuracy, it achieved an average Minimum Distance to Target of 13.00 mm, outperforming the conventional method's 17.00 mm for an average improvement of 4.00 mm. For efficiency, it required only 20.2 Steps to Minimum Distance, significantly fewer than the 31.8 steps needed by the conventional method.

C. Analysis of Attention Behavior

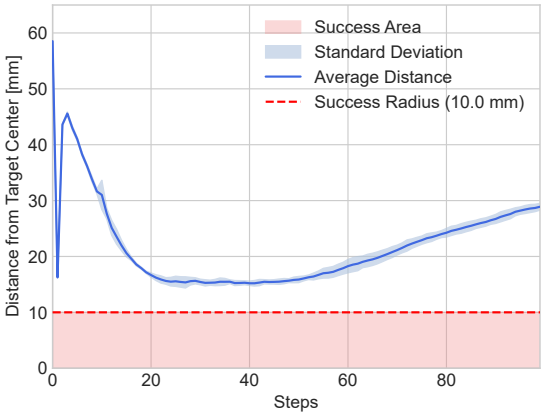
The performance difference stems from the quality of attention allocation. Fig. 7 shows a snapshot of the attention points (blue markers) for both models during autonomous control. In the conventional approach (a), attention often scatters to visually prominent but task-irrelevant areas, such as the shadow of the forceps. In contrast, the proposed method (b) maintains stable focus on the task-relevant objects: the target and the forceps themselves.

This difference becomes clearer when comparing the raw Attention Maps shown in Fig. 8, where warmer colors (red) represent higher attention scores, while cooler colors (blue) represent lower scores. In the conventional method's map (Fig. 8 (a)), while Channels 1 and 2 successfully track the target, Channels 3 through 6 exhibit strong responses to background areas and lighting reflections—features completely unrelated to the task. This tendency for attention to

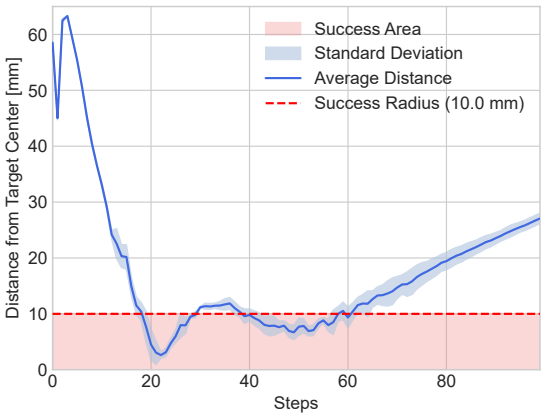
TABLE I
PERFORMANCE ON THE APPROACH TASK (MEAN ± SD)

Model	Metric	I	II	III	IV	V	Overall
Conv.	Minimum Distance to Target [mm]	11.16 ± 1.23	14.92 ± 1.54	19.94 ± 2.15	22.45 ± 2.51	16.55 ± 1.82	17.00 ± 3.10
	Steps to Minimum Distance	33.0 ± 3.0	35.0 ± 2.0	30.0 ± 4.0	25.0 ± 2.0	36.0 ± 3.0	31.8 ± 4.2
Prop.	Minimum Distance to Target [mm]	**10.00 ± 0.51**	**10.00 ± 0.62**	**16.75 ± 1.45**	**18.24 ± 1.78**	**10.00 ± 0.73**	**13.00 ± 2.50**
	Steps to Minimum Distance	**21.0 ± 2.0**	**20.0 ± 1.0**	**20.0 ± 2.0**	**20.0 ± 1.0**	**20.0 ± 2.0**	**20.2 ± 0.9**

Note: Values represent mean ± standard deviation (SD) for n=3 trials.



(a) Conventional Model (SARNN)



(b) Proposed Model (TAIRNN)

Fig. 6. Comparison of distance-to-target trajectories for the conventional model (a) and the proposed model (b) in a representative trial from initial position pattern II. The shaded region indicates the success area (distance < 10 mm). While the conventional model (a) fails to consistently stay within this area, the proposed model (b) successfully enters the region at approximately 20 steps and maintains proximity to the target thereafter.

be diverted toward task-irrelevant features likely contributes to the control instability.

By contrast, the proposed method's attention map (Fig. 8 (b)) shows most channels consistently directing attention to task-critical regions containing both the target object and the forceps, without the attention straying to task-unrelated background areas as seen in the conventional approach.

V. Discussion

As shown in Table I and Figs. 5 and 6, the proposed method outperforms the conventional approach in both accuracy and efficiency. The fundamental reason for this advantage is the qualitative difference in attention allocation between the models.

Figs. 7 and 8 clearly demonstrate this distinction. The conventional attention mechanism, which does not consider task context, tends to respond to visually prominent patterns within the image. As a result, as shown in Fig. 8 (a), its attention scatters not only on the target objects (Ch. 1,2) but also on irrelevant features like background elements and reflections (Ch. 3-6). This explains the unstable attention behavior observed in Fig. 7 (a) and the resulting poor final control performance.

In contrast, the proposed attention mechanism uses queries generated from the robot state a_t to enable context-aware attention. As shown in Fig. 8 (b), its focus remains consistently on task-relevant objects—the target and the forceps—while ignoring background noise. This ability to reliably track essential objects provides the LSTM with high-quality temporal information, which in turn enables robust and accurate motion generation.

However, as shown in Table I, both models consistently failed in patterns III and IV, which correspond to cases where the target object was located on the right. This suggests a fundamental limitation stemming from a combination of data bias and physical task constraints.

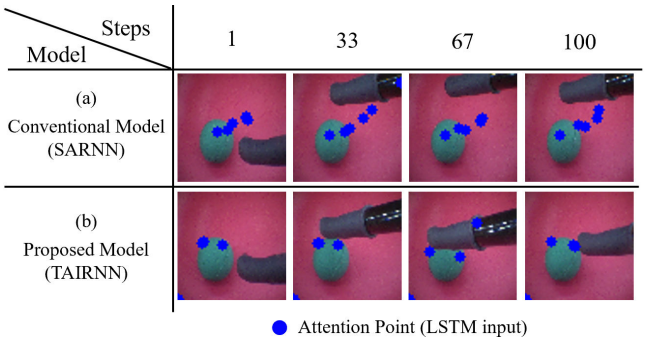
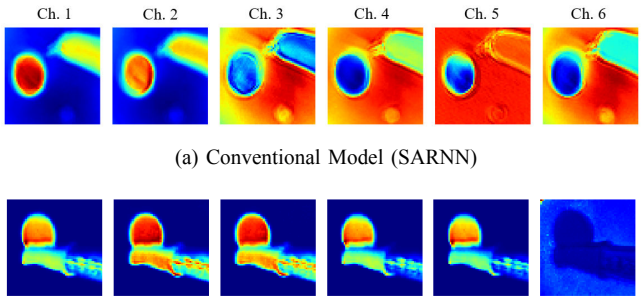


Fig. 7. Time-series comparison of attention points (blue markers). (a) The conventional model's attention is unstable and frequently distracted by task-irrelevant background noise like shadows. (b) The proposed model's attention remains focused on the task-relevant area of the tool and target, ignoring background noise.



(a) Conventional Model (SARNN)

(b) Proposed Model (TAIRNN)

Fig. 8. (a) The conventional model shows inconsistent behavior; while some channels (Ch. 1, 2) find the target, others (Ch. 3-6) are distracted by task-irrelevant background features. (b) The proposed model consistently focuses its attention on the task-relevant region containing both the target and the tool. Note that most channels respond to both objects.

Analysis of successful trials reveals that both models learned a strong bias to approach the target with the forceps from the right side (Fig. 5). This is likely because the training data was dominated by demonstrations using the right-hand manipulator. Consequently, the models learned this "right-side approach" as the optimal strategy.

In patterns Ⅲ and Ⅳ, where the target is already on the right, the models attempt to execute this biased strategy by maneuvering even further to the right. This action, however, leads to a critical failure condition: the forceps moves partially or fully out of the fixed endoscope's field of view. The resulting loss of visual feedback causes the attention mechanism to malfunction, preventing the LSTM from generating a corrective motion.

Therefore, the failures in patterns Ⅲ and Ⅳ are attributable to the interplay between an inappropriate policy learned from biased data and the physical limitations of a fixed-camera setup.

VI. CONCLUSIONS

This paper proposed a time series prediction model called TAIRNN (Task-Attentive Informed Recurrent Neural Network) that incorporates a cross-attention mechanism into the framework of deep predictive learning, which dynamically controls the target of attention using the robot's own state. Experiments on a surgical robot demonstrated that our method outperforms conventional approaches that rely solely on image information. The proposed method improved the Minimum Distance to Target by an average of 4.0 mm, enhancing both accuracy and efficiency.

In this study, we observed failures in scenarios that were not represented in the training data, where the surgical instrument moved beyond the fixed camera's field of view and visual feedback was lost. To address these limitations, future work will diversify the dataset to cover a wider range of initial conditions and manipulation phases. This includes continuous sequences that span approach−grasp−transport−placement and scenes with stronger visual disturbances frequently observed in endoscopic surgery. We also plan to include trajectories in which the tool temporarily exits and then re-enters the camera's view, enabling the model to learn recovery from interrupted visual input and to maintain robust attention under more realistic conditions.

References

[1] A. Mohan, U. U. Wara, M. T. A. Shaikh, R. M. Rahman, Z. A. Zaidi, and M. T. A. Shaikh, "Telesurgery and robotics: an improved and efficient era," *Cureus*, vol. 13, no. 3, 2021.

[2] P. Picozzi, U. Nocco, G. Puleo, C. Labate, and V. Cimolin, "Telemedicine and robotic surgery: a narrative review to analyze advantages, limitations and future developments," *Electronics*, vol. 13, no. 1, p. 124, 2023.

[3] S. Schmidgall, J. D. Opfermann, J. W. Kim, and A. Krieger, "Will your next surgeon be a robot? Autonomy and AI in robotic surgery," *Science Robotics*, vol. 10, no. 104, p. eadt0187, 2025.

[4] J. Ge, M. Kam, J. D. Opfermann, H. Saeidi, S. Leonard, L. J. Mady, M. J. Schnermann, and A. Krieger, "Autonomous System for Tumor Resection (ASTR) −Dual-Arm Robotic Midline Partial Glossectomy," *IEEE Robotics and Automation Letters*, vol. 9, no. 2, pp. 1166–1173, 2023.

[5] T. Mikada, T. Kanno, T. Kawase, T. Miyazaki, and K. Kawashima, "Suturing support by human cooperative robot control using deep learning," *IEEE Access*, vol. 8, pp. 167739–167746, 2020.

[6] H. Saeidi, J. D. Opfermann, M. Kam, S. Wei, S. Léonard, M. H. Hsieh, J. U. Kang, and A. Krieger, "Autonomous robotic laparoscopic surgery for intestinal anastomosis," *Science Robotics*, vol. 7, no. 62, p. eabj2908, 2022.

[7] A. Attanasio, B. Scaglioni, E. De Momi, P. Fiorini, and P. Valdastri, "Autonomy in surgical robotics," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 4, no. 1, pp. 651–679, 2021.

[8] T. Haidegger, "Autonomy for surgical robots: Concepts and paradigms," *IEEE Transactions on Medical Robotics and Bionics*, vol. 1, no. 2, pp. 65–76, 2019.

[9] Y. Ou, A. Soleymani, X. Li, and M. Tavakoli, "Autonomous blood suction for robot-assisted surgery: A sim-to-real reinforcement learning approach," *IEEE Robotics and Automation Letters*, vol. 9, no. 8, pp. 7246–7253, 2024.

[10] K. Kawaharazuka, K. Okada, and M. Inaba, "Robotic constrained imitation learning for the peg transfer task in fundamentals of laparoscopic surgery," in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, 2024, pp. 606–612.

[11] B. Li, R. Wei, J. Xu, B. Lu, C. H. Yee, C. F. Ng, P.-A. Heng, Q. Dou, and Y.-H. Liu, "3D perception based imitation learning under limited demonstration for laparoscope control in robotic surgery," in *Proc. Int. Conf. Robotics and Automation (ICRA)*, 2022, pp. 7664–7670.

[12] J. W. Kim, T. Z. Zhao, S. Schmidgall, A. Deguet, M. Kobilarov, C. Finn, and A. Krieger, "Surgical robot transformer (SRT): Imitation learning for surgical tasks," *arXiv preprint arXiv:2407.12998*, 2024.

[13] H. Ito, K. Yamamoto, H. Mori, and T. Ogata, "Efficient multitask learning with an embodied predictive model for door opening and entry with whole-body control," *Science Robotics*, vol. 7, no. 65, p. eaax8177, 2022.

[14] H. Ichiwara, H. Ito, K. Yamamoto, H. Mori, and T. Ogata, "Contact-rich manipulation of a flexible object based on deep predictive learning using vision and tactility," in *Proc. Int. Conf. Robotics and Automation (ICRA)*, 2022, pp. 5375–5381.

[15] M. Gualtieri and R. Platt, "Learning manipulation skills via hierarchical spatial attention," *IEEE Trans. Robotics*, vol. 36, no. 4, pp. 1067–1078, 2020.

[16] G. Seneviratne, K. Weerakoon, M. Elnoor, V. Rajgopal, H. Varatharajan, M. K. M. Jaffar, J. Pusey, and D. Manocha, "CROSS-GAiT: Cross-attention-based multimodal representation fusion for parametric gait adaptation in complex terrains," *arXiv preprint arXiv:2409.17262*, 2024.

[17] J. Zhao, H. Chen, Q. Tian, J. Chen, B. Yang, Z. Zhang, and H. Liu, "BronchoCopilot: Towards autonomous robotic bronchoscopy via multimodal reinforcement learning," in *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS)*, 2024, pp. 6923–6930.

[18] K. Iwatani, F. Urabe, S. Saito, S. Kawano, T. Yamasaki, S. Kimura, H. Otsuki, K. Fujio, T. Kimura, and J. Miki, "Initial experience of a novel surgical assist robot 'Saroa' featuring tactile feedback and a roll-clutch system in radical prostatectomy," *Scientific Reports*, vol. 14, no. 1, p. 31727, 2024.