Deep Reinforcement Learning based Control Framework for Multilateral Telesurgery

Sarah Chams Bacha*, Weibang Bai*, *Member, IEEE,* Ziwei Wang*, *Member, IEEE,* Bo Xiao, *Member, IEEE,* and Eric M. Yeatman, *Fellow, IEEE*

***Abstract*—The upper boundary of time delay is often required in traditional telesurgery control design, which would result in infeasibility of telesurgery across regions. To overcome this issue, this paper introduces a new control framework based on deep deterministic policy gradient (DDPG) reinforcement learning (RL) algorithm. The developed framework effectively overcomes the phase difference and data loss caused by time delays, which facilitates the restoration of surgeon's intention and interactive force. Kalman filter (KF) is employed to blend multiple surgeons' commands and predict the final local commands, respectively. The control framework ensures synchronization tracking performance and transparency. Prior knowledge of time delay is therefore not required. Simulation and experiment results have demonstrated the merits of the proposed framework.**

***Index Terms*—Telesurgery, deep deterministic policy gradient, reinforcement learning, time delay.**

I. Introduction

TELEOPERATION provides surgeons with the ability to operate on the patient side remotely, where the human intention can be transferred through communication channels [1]–[4]. Teleoperation mechanism isolates surgeons from the patient's neighbouring, which fulfills the medical requirements for diagnosis and treatment of infectious diseases [5] and surgical operations [6].

Time delay induced by distanced communication is a critical issue in the control design of telesurgery. Reported literature focused on the time-delay stability of teleoperation, which can be mainly categorized into wave-variable-based, predictive-based and adaptive-based control (refer to the classic reviews [7]–[9]). In specific, adaptive control based on universal approximation properties has shown great potential to handle teleoperation systems with external disturbances [10]. Haptic feedback with application to eye surgery was discussed in [11], where a passivity-based approach with impedance parameter optimization was used. It has been shown that time delay in the communication channel is ideally less than 180ms [12] in terms of realizing haptic communication. For cross-regional teleoperation, prescribed-time control is possible to overcome the large latency issue [13], [14]. However, prior knowledge of time delays and their derivatives required in the aforementioned works is difficult to obtain in advance. Thus, overcoming large latency without prior information is still an open challenge.

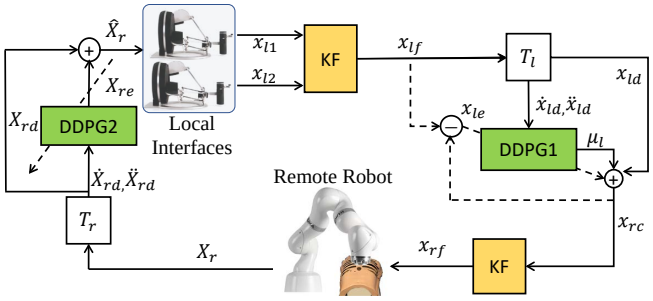

Fig. 1. Dual DDPG-based multilateral telesurgery framework diagram (The time step and the argument of variables are omitted).

Additionally, traditional telesurgery relies on the surgeon's full control of the robot on the patient's side, which would result in the surgeon's fatigue and degrade the surgery performance. To overcome this problem, introducing multiple surgeons with separate interfaces could reduce subjective operation errors effectively and therefore reduce the physical and mental workload of each surgeon. The normalisation technique was used to assign the authority weights in dual-user haptic systems [15], which was further applied in the subsequent teleoperation works [16], [17]. However, the above authority factors are pre-determined as constants such that the incorrect commands under large weights cannot be compensated for, which would weaken the coordination role of the multiple operators.

In this paper, we introduce a new deep RL based control framework for multilateral telesurgery, where dual KFs are utilized to blend surgeons' commands online and predict the final local signal on the remote side. The DDPG algorithm is employed to generate the optimized policy, which corrects the phase difference and data loss caused by unknown latency. The contributions are twofold: 1) The developed control framework enables the synchronization tracking and transparency performance simultaneously for telesurgery systems; 2) Prior knowledge of latency is not necessary in the multilateral collaboration scenarios.

II. Control Design

The overall control framework is depicted in Fig. 1, where a DDPG-based agent and a KF module are configured on local and remote side, respectively. The KF module on the local side fuses the multiple surgeons' commands to generate smooth

* These authors contributed equally. This research was supported by the European Commission Grants H2020 PH-CODING (FETOPEN 829186), CONBOTS (ICT 871803), and the UK EPSRC Grants (EP/P012779/1, EP/R026092/1). (Corresponding author: Ziwei Wang)

S. Chams Bacha and Ziwei Wang are with the Department of Bioengineering, Imperial College London, W12 0BZ, United Kingdom (e-mail: ziwei.wang@imperial.ac.uk).

Weibang Bai and Bo Xiao are with the Department of Computing, Imperial College London, SW7 2AZ, United Kingdom.

E. M. Yeatman is with Department of Electrical and Electronic Engineering, Imperial College London, SW7 2AZ, United Kingdom.

trajectories, while the KF on the remote side aims at filtering the final command for the remote robot. DDPG-based agent facilitates the reduction of phase difference caused by time delay. It is noted that the variables in Fig. 1 are defined in Cartesian space.

Surgeons' position commands (i.e., x_{l1} and x_{l2}) are firstly fused in the KF module with the blended output (x_{lf}). A DDPG-based agent (DDPG1) is then trained offline to add the specific instantaneous bias into the delayed signal (x_{ld}) depending on its velocity and acceleration, namely $\dot{x}_{ld}$ and $\ddot{x}_{ld}$, in the presence of forward time delay T_l. The optimization of DDPG1 is guided through the properly designed reward function, which minimizes the instantaneous signal error x_{le}. The resulting signal x_{rc} is then filtered through the second KF module, which helps overcome the adverse effect caused by the time step output in the neural networks. Furthermore, a state feedback controller such as PID controller determines the control torque for the remote robot, which updates the end-effector state and interactive force (grouped by $X_r = [x_r\ f_r]$) of the remote robot online. The other DDPG-based agent (DDPG2) is developed to enhance the transparency for the overall teleoperation system, thereby providing operators with estimated interactive forces and end-effector states in the presence of backward time delay (T_r).

A. Kalman Filter Module

The KF modules at the local and remote side assume the linear process system as:

$$\xi_{ijk} = A_i\xi_{ij(k-1)} + B_i u_{ijk} + \omega_{ik},\ \omega_{ik} \in \mathcal{N}(0, W_{ik}), \quad (1)$$

$$x_{ijk} = C_{ij}\xi_{ijk} + v_{ijk},\ v_{ijk} \in \mathcal{N}(0, R^j_{ik}), \quad (2)$$

where $k \in \mathbb{N}_+$ is the sample-step index with $\mathbb{N}_+$ being the positive integers. The subscript $i \in \{l, r\}$ stands for the local and remote side, respectively. The superscript $j \in \{1, 2\}$ denotes operators 1 and 2 on the local side, which only works for $i = l$. $\xi_{ljk} \triangleq [x_{ljk}, \dot{x}_{ljk}]$ and $\xi_{rk} \triangleq [x_{rck}, \dot{x}_{rck}]$. u_{ijk} is the acceleration vector, ω_{ik} the process noise with covariance $W_{ik} = \mathbb{E}[\omega_{ik}\omega_{ik}^{\mathrm{T}}]$, A_i the state transition matrix, B_i the control matrix, x_{ijk} the observed state, C_{ij} the projection matrix, and v_{ijk} the observation noise with covariance $R^j_{ik} = \mathbb{E}[v_{ijk}v_{ijk}^{\mathrm{T}}]$. At each sample, the state estimation and correction are derived by:

$$\xi^*_{ijk} = A_i\hat{\xi}_{ij(k-1)} + B_i u_{ijk}, \quad (3)$$

$$\hat{\xi}_{ijk} = \xi^*_{ijk} + K^j_{ik}(x_{ijk} - C_{ij}\xi^*_{ijk}), \quad (4)$$

where ξ^*_{ijk} is the state estimation, $\hat{\xi}_{ijk}$ is the corrected state, and K^j_{ik} is the Kalman gain expressed as:

$$K^j_{ik} = P^j_{i(k-1)}C_{ij}^{\mathrm{T}}[C_{ij}P^j_{i(k-1)}C_{ij}^{\mathrm{T}} + R^j_{ik}]^{-1}, \quad (5)$$

$$P^j_{ik} = A_i(P^j_{i(k-1)} - K^j_{ik}C_{ij}P^j_{i(k-1)})A_i^{\mathrm{T}} + W_{ik}. \quad (6)$$

In particular, the blended local command x_{lf} is obtained via track-to-track KF fusion such that

$$x_{lf} = x_{l1k} + (P^1_{lk} - P^{12}_{lk})(P^1_{lk} + P^2_{lk} - P^{12}_{lk} - P^{21}_{lk})^{-1}(x_{l2k} - x_{l1k}) \quad (7)$$

where $P^{12}_{lk} = (P^{21}_{lk})^{\mathrm{T}}$ is the cross covariance matrix between x_{l1k} and x_{l2k}, which is given by:

$$\begin{aligned} P^{12}_{lk} =& (I - K^1_{lk}C_{l1})A_l P^{12}_{l(k-1)}A_l^{\mathrm{T}}(I - K^2_{lk}C_{l2})^{\mathrm{T}} \\ &+ (I - K^1_{lk}C_{l1})W_{l(k-1)}(I - K^2_{lk}C_{l2})^{\mathrm{T}} \end{aligned} \quad (8)$$

where I is the identity matrix.

B. DDPG based Latency Compensator

In order to compensate for the latency, we adopt the same DDPG structure for two agents at local and remote side, which is composed of two neural networks, namely the deep Q network (critic network) that estimates the value function (9), and the policy network (actor network) that generates the optimal action

$$\hat{Q}_i(s_{ik}, a_{ik}) = \sum_{j=1}^{N_{iQ}} \phi^j_{ik}\sigma_{icj}(s_{ik}, a_{ik}), \hat{\mu}_i(s_{ik}) = \sum_{j=1}^{N_{i\mu}} \theta^j_{ik}\sigma_{iaj}(s_{ik}), \quad (9)$$

where $\hat{Q}_i(s_{ik}, a_{ik})$ is the estimation of value function $Q_i(s_{ik}, a_{ik})$ for state s_{ik} and action $a_{ik} = \mu_i(s_{ik})$ at the kth sample. $\hat{\mu}_i(s_{ik})$ is the estimation of optimal policy for s_{ik}. N_{iQ} and $N_{i\mu}$ are the number of basis functions in the output layer. ϕ^j_{ik} and θ^j_{ik} are the j-th network weights for the output layer. σ_{icj} and σ_{iaj} are the basis functions for the critic and actor networks, respectively. It is noted that the argument of the variables will be omitted without specific emphasis for simplicity. A dense model is used as the basis function while the ReLu is chosen as the activation function for both networks. Given the set of the whole weight parameters of the critic network as ϕ_i, the following cost function can be derived:

$$J(\phi_i) = \frac{1}{N_i}\sum_{k=1}^{N_i}(y_{ik} - Q_i(s_{ik}, a_{ik}))^{\mathrm{T}}(y_{ik} - Q_i(s_{ik}, a_{ik})), \quad (10)$$

and

$$y_{ik} = r_{ik} + \gamma_i Q^t_i(s_{i(k+1)}, \mu^t_i(s_{i(k+1)})), \quad (11)$$

where N_i is the number of transitions in the mini-batch experience, r_{ik} is the immediate reward, and $\gamma_i \in (0, 1)$ is the discount factor. Q^t_i and μ^t_i are the target value and policy. With the gradient descent method, one can obtain the parameters updated by minimizing (10)

$$\phi_i \leftarrow \phi_i - \alpha^Q_i \nabla_{\phi_i} J(\phi_i) \quad (12)$$

where α^Q_i is the learning rate of the critic network. With the same mini-batch experience to search for the optimal policy, the actor network updates the parameters by approximating the deterministic policy gradient [18]:

$$\nabla_{\theta_i} Q_i(s_i, a_i) \approx \frac{1}{N_i}\sum_{k=1}^{N_i} \nabla_{a_{ik}} Q_i(s_{ik}, a_{ik})\nabla_{\theta_i}\mu_i(s_{ik}) \quad (13)$$

where the set of the whole weight parameters of the actor network is denoted by θ_i. Let α^μ_i be the learning rate of the actor network, the weight parameters of the actor network is therefore updated by

$$\theta_i \leftarrow \theta_i - \alpha^\mu_i \nabla_{\theta_i} Q_i(s_i, a_i). \quad (14)$$

Given that $\tau \ll 1$ is a positive constant, combining (12) and (14) yields the weight parameters of the target critic and actor networks softly updated by

$$\phi_i^t \leftarrow \tau\phi_i + (1-\tau)\phi_i^t, \qquad \theta_i^t \leftarrow \tau\theta_i + (1-\tau)\theta_i^t. \quad (15)$$

The reward function is defined as:

$$r_{ik}^{\nu} = \begin{cases} -r_{i1}X_{ie_k}^{\nu} & \text{if } X_{ie_k}^{\nu} > \epsilon_{i1} \\ r_{i2} & \text{if } \epsilon_{i1} \le X_{ie_k}^{\nu} \le \epsilon_{i2} \\ r_{i3} & \text{else} \end{cases} \quad (16)$$

where the superscript $\nu \in \mathbb{N}_+$ is the ν-th entry of the corresponding vector. $X_{le_k}^{\nu} \triangleq x_{le_k}^{\nu} = |x_{rc_k}^{\nu} - x_{lf_k}^{\nu}|$, x_{rc} is the resulting signal on the remote side, and x_{lf} is the filtered signal generated on the local side. Similarly, $X_{re_k}^{\nu} \triangleq |\hat{X}_{r_k}^{\nu} - X_{r_k}^{\nu}|$, in which $\hat{X}_r$ includes the post-compensation state and force feedback by DDPG2. ϵ_{i1} and ϵ_{i2} are positive thresholds to evaluate the tracking error. The reward parameters $r_{i1} > 0$, $0 < r_{i2} < r_{i3}$ can be predefined according to tasks.

Remark 1. The reward parameters and thresholds in (16) can be user-defined according to synchronization tracking and transparency requirements. ϵ_{i1} and ϵ_{i2} in the judging condition represent the desirable accuracy in transient/steady-state process, respectively, which are subject to measurement accuracy. The reward parameters depend on the demand of surgery requirements. In general, large r_{i1} and r_{i3} are required to compensate for the phase difference.

Remark 2. The following procedures are summarized when implementing the proposed dual DDPG algorithm in telesurgery: (i) determine the KF system matrices, reward parameters and thresholds according to Remark 1, (ii) collect the surgeon's motor data (x_{lf}) and remote interaction information (X_r), (iii) develop actor and critic networks and update the network parameters as (9)-(15) on local and remote sides, respectively.

III. Simulation and Experiment

The control framework is firstly tested through simulation in comparison with [19]. The remote manipulator is chosen as a SCARA planar robot, while the local commands are generated by dual 7-DoF Omega.7 interfaces (Force Dimension Inc., Switzerland). The parameters of both KFs are empirically set as follows: $A_i = \begin{bmatrix} 1 & 10^{-3} \\ 0 & 1 \end{bmatrix}$, $B_i = \begin{bmatrix} 10^{-6}/2 & 0 \\ 0 & 10^{-6}/2 \end{bmatrix}$, $W_{i0} = \begin{bmatrix} 10^{-12}/36 & 10^{-10}/12 \\ 10^{-10}/12 & 10^{-8}/4 \end{bmatrix}$, $P_{i0}^j = \begin{bmatrix} 10^{15} & 0 \\ 0 & 10^{15} \end{bmatrix}$, $C_{ij} = [1, 0]$, $R_{i0}^j = 10^{-6}$, $\xi_{i10} = [0, -0.21]^{\mathrm{T}}$, $\xi_{i20} = [0, 0.09]^{\mathrm{T}}$ for $i \in \{l, r\}$ and $j \in \{1, 2\}$. To guarantee the safety requirements of the experimental setup, we employ the local regression using weighted linear least squares and polynomial model before implementing x_{rf}. Similarly with [19], [20], we select the tissue damping and stiffness coefficients as $D = 10$N·s/m and $K = 250$N/m for remote palpation, which is a key step in clinical surgery. The critic network is composed of three hidden layers with five neurons each, and one output layer with two neurons, while the actor network one hidden layer with three neurons. Adaptive Moment Estimation (ADAM) is employed with the learning rate $\alpha_i^Q = 0.01$ and $\alpha_i^\mu = 0.01$. The discount factor is set as 0.99, mini-batch size 100, noise variance 0.01, and variance decay rate 10^{-5}. The RL agents are trained using the generated trajectories for 20 episodes and each episode consisting of 2000 steps, where each step size is 0.01s. A boundary of $[-1, 1]$ is set for the output of the RL module. The agent is trained under random time delay in the range of $[0.1, 1]$s. $\epsilon_{i1} = 0.1$, $\epsilon_{i2} = 0.01$, $r_{i1} = r_{i3} = 10$, $r_{i2} = 2$ for $i \in \{l, r\}$.

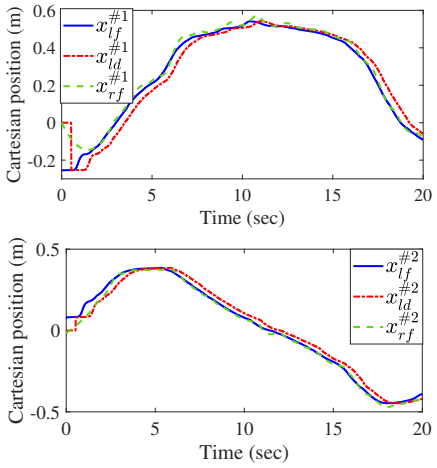

Fig. 2. Cartesian position tracking profiles.

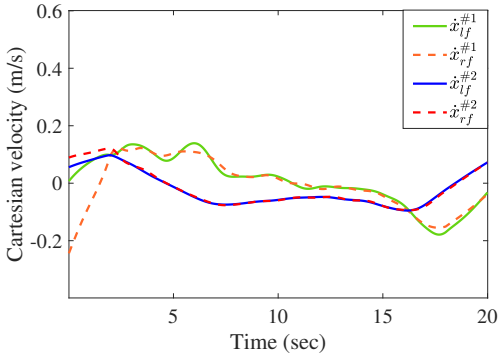

Fig. 3. Cartesian velocity tracking profiles.

The simulation results in the case of $T_l = 0.5$s are shown in Figs. 2 and 3. We use the normalised root-mean-squared error (NRMSE) to evaluate the degree of local intention restoration quantitatively, with the x_{rc}-x_{lf} pair as example:

$$\mathrm{NRMSE}(x_{rc}^{\#p}) = \frac{\sqrt{\frac{1}{N}\sum_{k=1}^{N}(x_{rck}^{\#p} - x_{lfk}^{\#p})^2}}{\max(x_{lfk}^{\#p}) - \min(x_{lfk}^{\#p})} \quad (17)$$

where $N = 2000$. $p \in \{1, 2\}$ denotes the x- and y-direction, respectively. Comparative simulation results are summarized in Tab. I, where the pairs x_{rc}-x_{lf}, $\dot{x}_{rc}$-$\dot{x}_{lf}$, and $\hat{f}_r$-f_r are tested in three latency cases, namely $T_1 = (0.3 + \delta_1)$s, $T_2 = (0.6 + \delta_2)$s, $T_3 = (0.8 + \delta_3)$s, $\delta_1 \sim N(0.1, 0.004)$, $\delta_2 \sim N(0.2, 0.006)$, and $\delta_3 \sim N(0.2, 0.006)$.

We have also carried out experiments to validate the proposed control framework. A KUKA iiwa 14 robot is used as

the remote side controlled by dual 7-DoF Omega. 7 interfaces via local area network (LAN). An additional 0.5s delay is added to test the control performance. The filtered signal x_{lf} is sent to the remote robot through the Robot Operating System (ROS). The control period for each loop is set as 5ms. The experimental results including the end-effector position and velocity are directly collected from the controller in real time, as shown in Figs. 4 and 5.

TABLE I
PERFORMANCE METRICS UNDER BILATERAL CONTROLLER [19] AND THE PROPOSED DUAL-DDPG (D-D) FRAMEWORK IN THREE CASES OF TIME DELAYS.

T_l,T_r / NRMSE	T_1		T_2		T_3	
	[19]	D-D	[19]	D-D	[19]	D-D
$x_{rc}^{\#1}$(m)	0.064	**0.042**	0.085	**0.068**	0.103	**0.073**
$x_{rc}^{\#2}$(m)	0.036	**0.033**	0.064	**0.061**	0.077	**0.056**
$\dot{x}_{rc}^{\#1}$(m/s)	0.132	**0.095**	0.294	**0.081**	0.292	**0.098**
$\dot{x}_{rc}^{\#2}$(m/s)	0.131	**0.083**	0.342	**0.090**	0.254	**0.091**
$\hat{f}_r^{\#1}$(N)	0.064	**0.044**	0.108	**0.087**	0.127	**0.105**
$\hat{f}_r^{\#2}$(N)	0.564	**0.036**	0.586	**0.062**	0.597	**0.086**

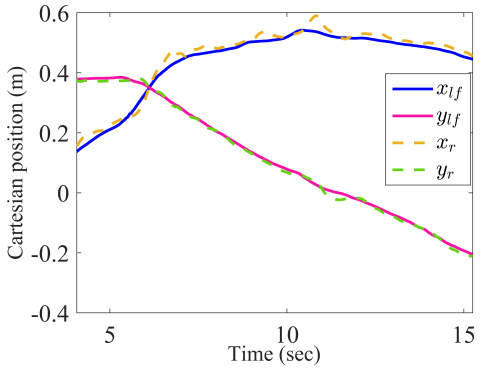

Fig. 4. Cartesian position tracking profiles in experiment.

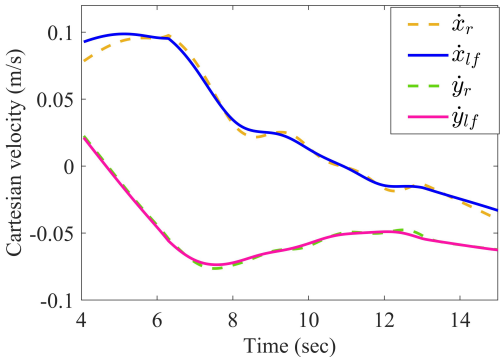

Fig. 5. Cartesian velocity tracking profiles in experiment.

It can be observed from Tab. I that, compared with traditional bilateral control [19], the accurate restoration of the less-delayed human command signal is realized with significant decrease in the instantaneous error. Fig. 4 shows the effectiveness of the proposed approach in dealing with time delays, where the end-effector of the remote robot can follow the local position command in x and y coordinate without phase difference. The corresponding RMSEs are 0.0192m and 0.0086m in terms of Cartesian position tracking. The velocity tracking profile is depicted in Fig. 5, where the RMSEs of the tracking errors in x and y coordinate are 0.0042m/s and 0.0014m/s, respectively.

IV. Conclusion

In this paper, a new control framework has been presented for multilateral telesurgery. Based on Kalman filter and the DDPG algorithm, the human command and interactive force can effectively be transferred between the local and remote side without phase difference, thereby eliminating the adverse effect caused by unknown time delay. The prior knowledge of time delay is not necessary to be known. The proposed framework allows for multilateral collaboration to blend subjective decisions, which enables the synchronization tracking and transparency performance simultaneously. User assessment and in-vivo telesurgery experiments with external disturbances will be implemented in the future.

References

[1] C. Bergeles and G.-Z. Yang, "From passive tool holders to microsurgeons: safer, smaller, smarter surgical robots," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 5, pp. 1565–1576, 2013.
[2] J. Troccaz, G. Dagnino, and G.-Z. Yang, "Frontiers of medical robotics: from concept to systems to clinical translation," *Annu. Rev. Biomed. Eng.*, vol. 21, pp. 193–218, 2019.
[3] W. Bai, Q. Cao, P. Wang, P. Chen, C. Leng, and T. Pan, "Modular design of a teleoperated robotic control system for laparoscopic minimally invasive surgery based on ros and rt-middleware," *Ind. Robot*, vol. 44, no. 55, pp. 596–608, 2017.
[4] W. Bai, Q. Cao, C. Leng, Y. Cao, M. G. Fujie, and T. Pan, "A novel optimal coordinated control strategy for the updated robot system for single port surgery," *Int. J. Med. Robot. Comput. Assist. Surg.*, vol. 13, no. 3, p. e1844, 2017.
[5] A. Gao, R. R. Murphy, W. Chen, G. Dagnino, P. Fischer, M. G. Gutierrez, D. Kundrat, B. J. Nelson, N. Shamsudhin, H. Su, J. Xia, A. Zemmar, D. Zhang, C. Wang, and G.-Z. Yang, "Progress in robotics for combating infectious diseases," *Science Robotics*, vol. 6, no. 52, 2021.
[6] W. Bai, Z. Wang, Q. Cao, H. Yokoi, M. G. Fujie, E. M. Yeatman, and G.-Z. Yang, "Anthropomorphic dual-arm coordinated control for a single-port surgical robot based on dual-step optimization," *IEEE Trans. Med. Robot. Bionics*, vol. 4, no. 1, pp. 72–84, 2022.
[7] D. Sun, F. Naghdy, and H. Du, "Application of wave-variable control to bilateral teleoperation systems: A survey," *Annu. Rev. Control*, vol. 38, no. 1, pp. 12–31, 2014.
[8] R. Uddin and J. Ryu, "Predictive control approaches for bilateral teleoperation," *Annu. Rev. Control*, vol. 42, pp. 82–99, 2016.
[9] L. Chan, F. Naghdy, and D. Stirling, "Application of adaptive controllers in teleoperation systems: A survey," *IEEE Trans. Hum. Mach. Syst.*, vol. 44, no. 3, pp. 337–352, 2014.
[10] Z. Chen, F. Huang, C. Yang, and B. Yao, "Adaptive fuzzy backstepping control for stable nonlinear bilateral teleoperation manipulators with enhanced transparency performance," *IEEE Trans. Ind. Electron.*, vol. 67, no. 1, pp. 746–756, 2019.
[11] A. Mahdizadeh, M. A. Nasseri, and A. Knoll, "Transparency optimized interaction in telesurgery devices via time-delayed communications," in *2014 IEEE Haptics Symposium (HAPTICS)*, 2014, pp. 603–608.
[12] E. Ivanova, J. Eden, S. Zhu, G. Carboni, A. Yurkewich, and E. Burdet, "Short time delay does not hinder haptic communication benefits," *IEEE Trans. Haptics*, vol. 14, no. 2, pp. 322–327, 2021.
[13] Z. Wang, B. Liang, Y. Sun, and T. Zhang, "Adaptive fault-tolerant prescribed-time control for teleoperation systems with position error constraints," *IEEE Trans. Ind. Informat.*, vol. 16, no. 7, pp. 4889–4899, 2020.

[14] Z. Wang, H. Lam, B. Xiao, Z. Chen, B. Liang, and T. Zhang, "Event-triggered prescribed-time fuzzy control for space teleoperation systems subject to multiple constraints and uncertainties," *IEEE Trans. Fuzzy Syst.*, vol. 29, no. 9, pp. 2785–2797, 2021.

[15] B. Khademian and K. Hashtrudi-Zaad, "Dual-user teleoperation systems: New multilateral shared control architecture and kinesthetic performance measures," *IEEE/ASME Trans. Mechatronics*, vol. 17, no. 5, pp. 895–906, 2012.

[16] M. Shahbazi, S. F. Atashzar, H. A. Talebi, and R. V. Patel, "Novel cooperative teleoperation framework: Multi-master/single-slave system," *IEEE/ASME Trans. Mechatronics*, vol. 20, no. 4, pp. 1668–1679, 2015.

[17] J. H. Ryu, Q. Ha-Van, and A. Jafari, "Multilateral teleoperation over communication time delay using the time-domain passivity approach," *IEEE Trans. Control Syst. Technol.*, vol. 28, no. 6, pp. 2705–2712, 2020.

[18] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.

[19] Z. Chen, B. Liang, T. Zhang, and X. Wang, "Bilateral teleoperation in Cartesian space with time-varying delay," *Int. J. Adv. Robot. Syst.*, vol. 9, no. 4, p. 110, 2012.

[20] H. Li and A. Song, "Virtual-environment modeling and correction for force-reflecting teleoperation with time delay," *IEEE Trans. Ind. Electron.*, vol. 54, no. 2, pp. 1227–1233, 2007.