*Robotica* (2023), **41**, pp. 1115–1126
doi:10.1017/S0263574722001527

RESEARCH ARTICLE

# Discrete soft actor-critic with auto-encoder on vascular robotic system

Hao Li[1,2], Xiao-Hu Zhou[1,2,*], Xiao-Liang Xie[1,2,*], Shi-Qi Liu[1,2], Mei-Jiang Gui[1,2], Tian-Yu Xiang[1,2], Jin-Li Wang[3] and Zeng-Guang Hou[1,2,4]

[1]The State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China, [2]The School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China, [3]The School of Mechanical Electronic and Information Engineering, China University of Mining and Technology, Beijing, China, and [4]The CAS Center for Excellence in Brain Science and Intelligence Technology, Beijing, China
*Corresponding authors. E-mail: xiaohu.zhou@ia.ac.cn; xiaoliang.xie@ia.ac.cn

**Received:** 19 May 2022; **Revised:** 29 September 2022; **Accepted:** 29 September 2022;
**First published online:** 17 November 2022

**Keywords:** surgical robots, vascular robotic system, automation, reinforcement learning, deep neural network

## Abstract

Instrument delivery is critical part in vascular intervention surgery. Due to the soft-body structure of instruments, the relationship between manipulation commands and instrument motion is non-linear, making instrument delivery challenging and time-consuming. Reinforcement learning has the potential to learn manipulation skills and automate instrument delivery with enhanced success rates and reduced workload of physicians. However, due to the sample inefficiency when using high-dimensional images, existing reinforcement learning algorithms are limited on realistic vascular robotic systems. To alleviate this problem, this paper proposes discrete soft actor-critic with auto-encoder (DSAC-AE) that augments SAC-discrete with an auxiliary reconstruction task. The algorithm is applied with distributed sample collection and parameter update in a robot-assisted preclinical environment. Experimental results indicate that guidewire delivery can be automatically implemented after 50k sampling steps in less than 15 h, demonstrating the proposed algorithm has the great potential to learn manipulation skill for vascular robotic systems.

## 1. Introduction

Vascular intervention surgery (VIS) is a mainstay for the treatment of coronary artery disease, which is a major cause of death worldwide [1, 2]. VIS is achieved through the use of catheters, guidewires, and other instruments that are percutaneously introduced into the vasculature and navigated to target locations within the vascular system [3, 4]. During VIS, physicians use X-ray fluoroscopy for guidance and are required to wear heavy lead-lined garments for radiation protection, causing radiation-associated hazards [5–7] and orthopedic strain injuries [8]. To reduce those risks, vascular robotic systems with the master-slave structure and tactile perception [9–12] have been developed. Clinical trials have shown the advantages of vascular robotic systems in X-ray exposure reduction, instrument control precision, procedural duration reduction, and remote operation [13].

Learning instrument-manipulation skills, which is the core in VIS, requires extensive specialized training for physicians. VIS instruments are designed as soft-body structures to advance in tortuous vessels. Due to soft-body structures of instruments, the relationship between manipulation commands and instrument motion is non-linear, making instrument delivery challenging and time-consuming. In addition, the diversity of VIS instruments and scenarios also greatly extends the training cycle for VIS. Vascular robotic systems are expected to have higher-level autonomy [14], assist physicians in some operations, reduce the difficulty of VIS, and shorten training time for VIS. Besides shorter training time, potential advantages of vascular robotic systems with high-level autonomy include shorter surgery time,

https://doi.org/10.1017/S0263574722001527 Published online by Cambridge University Press

increased accuracy of instrument manipulation [15], and reduced fatigue of clinicians. However, most existing vascular robotic systems have no autonomy and completely follow manual teleoperation.

Learning instrument-manipulation skills and augmenting vascular robotic systems with high-level autonomy has attracted a wide range of research interests. Statistical models [16–18] and imitation learning [19–21] have been used to automate instrument delivery. However, both statistical model and imitation learning require high-quality datasets. Due to the scarcity of VIS data and the diversity of VIS scenarios, it is different to obtain a large-scale comprehensive dataset covering most scenarios.

Reinforcement learning (RL) is achieved by trial-and-error without prior labeled datasets and has been widely used in many fields including games, robots, and autonomous vehicles [22–24]. Some research uses RL algorithms to learn instrument-manipulation skills based on vectorized low-dimensional representations of VIS scenarios [21, 25–27]. However, in clinical scenarios, it is difficult to obtain vectorized information such as the position and velocity of instruments. Analogous to physicians using X-ray fluoroscopy for intraoperative navigation [28], it is more realistic to use images for instrument-manipulation skills. Several state-of-the-art RL algorithms, such as Dueling Deep Q-Network [29] and Asynchronous Advantage Actor-Critic [30], have been applied with preoperative vascular models to learn instrument-manipulation skills with high-dimension images [31, 32]. Due to notorious sample inefficiency of RL, existing research is limited to digital simulation environments [31, 32]. However, there is a non-negligible gap between digital simulations and real environments, which limits the clinical value of learned instrument-manipulation skills. VIS instruments are designed as soft-body structures, which are difficult to model for digital simulations. Moreover, motion deviation, communication fluctuation, friction, and other disturbances are also challenging to simulate. Therefore, applying RL with realistic vascular robotic systems is of great value for more realistic instrument-manipulation skills. However, sample collection in real-world environments is much slower than in simulation. Furthermore, to apply RL algorithms in clinic, the training should be completed within the interval between obtaining the preoperative vascular model and performing VIS. Thus, RL algorithms with higher sample efficiency are required. Some unsupervised learning methods in computer vision can be used as auxiliary tasks to improve sample efficiency of RL [33, 34]. But to the best of our knowledge, such methods have not been used for instrument-manipulation skill learning.

The main contributions of this paper are as follows:

(1) This paper proposes a novel RL method DSAC-AE that uses an auxiliary reconstruction task to accelerate training with foreground-background imbalanced images.

(2) Sample collection and parameter update of DSAC-AE are executed distributedly to shorten training time with realistic vascular robotic systems.

(3) Preclinical experiments demonstrate that DSAC-AE has high sample efficiency and can learn guidewire delivery within 50k sampling steps (less than 15 h).

(4) Ablation experiments demonstrate the effectiveness of auxiliary loss function choice and reward function design.

The paper is structured as follows: Section 2 introduces the preclinical environment and the proposed algorithm. Sections 3 and 4 show and analyze the results, respectively. Section 5 summarizes the paper.

## 2. Material and method

### 2.1. Environment and problem definition

As shown in Fig. 1, the whole environment consists of a vascular robotic system, a 3D-printed vascular model, a guidewire, and a catheter. The vascular robotic system (hereinafter referred to as robot) has two degrees of freedom (translation and rotation) [35]. The robot needs to deliver the guidewire to a target in the vascular model. The target is illustrated in Fig. 1(c). The size of the vascular model is $15 \times 15 \times 1.5$ cm, where vessels are about $3-5$ mm thick. The environment is randomly reset to avoid overfitting to specific

https://doi.org/10.1017/S0263574722001527 Published online by Cambridge University Press

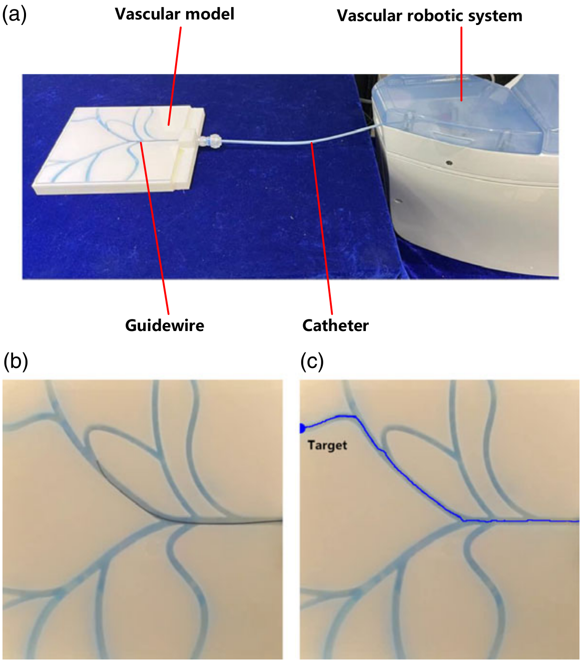

***Figure 1.*** *The preclinical environment. (a) The whole environment. The camera and lights are omitted for simplicity. (b) Close-up of the guidewire and the vascular model. (c) The target in the vascular model and the correct path from the outset to the target.*

states. Episodes start in initial states where the distal tip of the guidewire is randomly 10−15 pixels from the outset and the guidewire is rotated randomly, and end when the distal tip reaches the target or borders.

The guidewire delivery problem is defined as a partially observable Markov decision process $\langle \mathcal{S}, \mathcal{O}, \mathcal{A}, P, R, \gamma \rangle$, where $\mathcal{S}$ is the state space, $\mathcal{O}$ is the observation space, $\mathcal{A}$ is the action space, $P{:}\mathcal{S} \times \mathcal{A} \times \mathcal{S} \mapsto [0, 1]$ is the transition probability, $R{:}\mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$ is the reward function, and $\gamma$ is the discount factor. At time-step $t$, the robot receives an observation $o_t \in \mathcal{O}$ and chooses an action $a_t \in \mathcal{A}$ based on the stochastic policy $\pi(\cdot|o_t)$. Observations are $140 \times 140$ size images taken from a fixed camera directly above the vascular model. For generality and simplicity, each observation is prepossessed into binary images of the vessel and the guidewire as in Fig. 2. Actions are speed commands in translation and rotation freedom. The action space consists of 10 discrete actions. There are two sub-actions in the translation freedom, which are forward and backward with the same speed, and five sub-actions in the rotation freedom, including static, two-speed counterclockwise, and clockwise rotation. The interval between actions is 0.5 s. Then the state of environment $s_t$ changes to $s_{t+1}$ according to the transition probability $P(s_{t+1}|s_t, a_t)$, and the robot gets a reward $r_t = R(s_t, a_t)$ and a new observation $o_{t+1}$. The goal is to find optimal policy that maximizes the maximum entropy objective $\sum_t \mathbb{E}_{o_t}\{\mathbb{E}_{a_t\sim\pi(\cdot|o_t)}\{r_t + \alpha\mathcal{H}[\pi(\cdot|o_t)]\}\}$ [36], where $\alpha$ is the temperature parameter that balances the reward and the policy stochasticity. In the following, the subscript indicating the time-step will be truncated if the time-step does not need to be considered.

The reward $r_t$ guides the learning process and requires careful design. To ensure efficacy and safety, the reward consists of three sub-rewards :

(1) The sparse sub-reward that indicates completion of guidewire delivery. The agent will receive a large sparse bonus of size 400 if the target is reached; otherwise, the sparse sub-reward is 0. The guidewire is considered as reaching the target when the distal tip of the guidewire is within five pixels from the target.

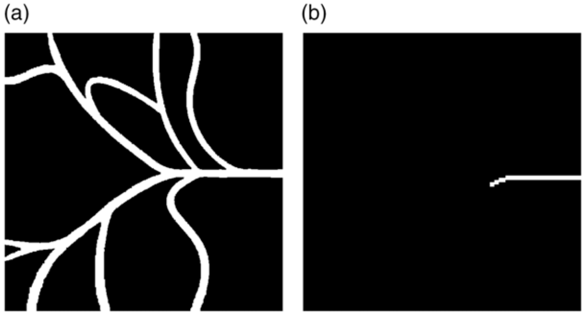

***Figure 2.*** *Prepossessed images as observations in RL. (a) The vessel. (b) The guidewire.*

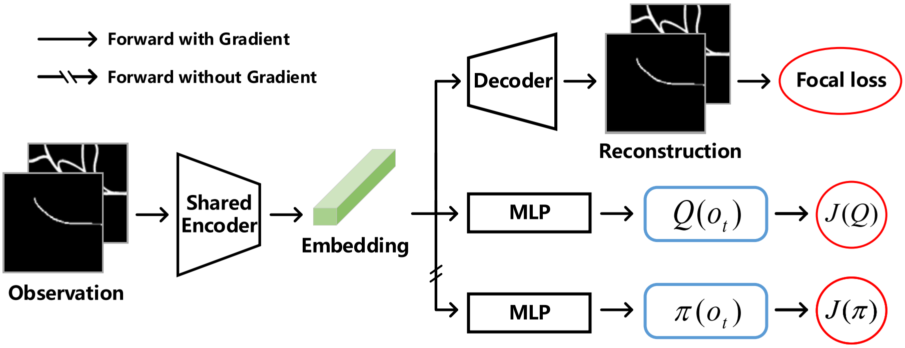

***Figure 3.*** *DSAC-AE: SAC-Discrete is augmented with an auto-encoder. The encoder is updated with gradients from soft Bellman residual $J(Q)$ and the auxiliary reconstruction task. Focal loss is used as the loss function of auxiliary reconstruction task to mitigate the impact of foreground-background imbalance.*

(2) The dense sub-reward that encourages the guidewire to approach the target along the correct path. The correct path is the shortest path from the outset to the target as shown in Fig. 1(c), which is automatically generated by the Dijkstra's algorithm. When the guidewire keeps on the correct path, the robot will get a reward that is equal to the decrease in distance to the target. The distance of each point to the target is calculated by the Dijkstra's algorithm simultaneously with the correct path. The robot will get a penalty of size 20 if the guidewire deviates from the correct path and goes into a wrong vessel branch, and a bonus of size 20 if the guidewire leaves the wrong vessel branch and goes back to the correct path. The robot will receive another penalty of size 50 if the guidewire exits the vascular model.

(3) The safe sub-reward. The robot will get a large penalty of size 200 if the contact force exceeds a safe threshold.

As it is a challenge to measure the contact force between the guidewire and the vascular model, safety during RL is ensured by limiting the contact force between the robot and the guidewire. The contact force between the robot and the guidewire is represented by the motor current.

### *2.2. Discrete soft actor-critic with auto-encoder*

To improve sample efficiency, we propose discrete soft actor-critic with auto-encoder (DSAC-AE) that uses an auxiliary reconstruction task as SAC-AE [33]. The whole architecture of DSAC-AE is shown in Fig. 3. The shared encoder compresses the input images into a vectorized embedding. Following the

https://doi.org/10.1017/S0263574722001527 Published online by Cambridge University Press

encoder, the decoder reconstructs the input, while the two multilayer perceptrons (MLP) represent soft Q-function $Q$ and policy $\pi$, respectively. The output of soft Q-function $Q$ and policy $\pi$ is $|\mathcal{A}|$-dimension vectors representing soft Q-value and probability of each action, respectively. In the following, $Q(o)$ and $\pi(o)$ represent output vectors of soft Q-function $Q$ and policy $\pi$ respectively, while $Q(o,a)$ and $\pi(a|o)$ represent elements corresponding to action $a$ in $Q(o)$ and $\pi(o)$, respectively.

Soft Q-function $Q$ and policy $\pi$ is updated according to SAC [37]. Soft Q-function $Q$ are trained by minimizing the soft Bellman residual

$$J(Q) = \mathbb{E}_{(o_t,a_t,r_t,o_{t+1})}\left\{\left[Q(o_t,a_t) - r_t - \gamma V(o_{t+1})\right]^2\right\}, \tag{1}$$

where $V$ is the soft state value function

$$V(o) = \mathbb{E}_{a\sim\pi(\cdot|o)}\left[\bar{Q}(o,a) - \alpha \log \pi(a|o)\right]. \tag{2}$$

$\bar{Q}$ denotes the target Q-function whose parameters are the exponentially moving average of soft Q-function parameters. Policy $\pi$ is updated by minimizing the following formula

$$J(\pi) = \mathbb{E}_o\left\{\mathbb{E}_{a\sim\pi(\cdot|o)}\{\alpha \log[\pi(a|o)] - Q(o,a)\}\right\}. \tag{3}$$

The temperature parameter $\alpha$ is updated by minimizing the following loss function

$$J(\alpha) = \mathbb{E}_o\left\{\mathbb{E}_{a\sim\pi(\cdot|o)}\{-\alpha \log[\pi(a|o)] - \alpha\bar{\mathcal{H}}\}\right\}, \tag{4}$$

where hyperparameter $\bar{\mathcal{H}}$ is the target entropy. Since the action space $\mathcal{A}$ is discrete, expectations on actions are directly computed without Monte Carlo method as in SAC-Discrete [38] to decrease variance of loss functions. This means that Eqs. (2), (3), and (4) are changed to:

$$V(o) = \pi(o)^T[\bar{Q}(o) - \alpha \log \pi(o)], \tag{5}$$

$$J(\pi) = \mathbb{E}_o\left\{\pi(o)^T\{\alpha \log[\pi(o)] - Q(o)\}\right\}, \tag{6}$$

$$J(\alpha) = \mathbb{E}_o\left\{-\alpha\pi(o)^T \log[\pi(o)] - \alpha\bar{\mathcal{H}}\right\}. \tag{7}$$

The decoder is updated by minimizing reconstruction error. Since vessels and guidewires are narrow, the foreground is much smaller than the background in prepossessed images. To mitigate the impact of foreground-background imbalance, focal loss [39] is used for the auxiliary reconstruction task:

$$\mathrm{FL}(\hat{o},o) = \sum_{i,j} -o_{ij}\left(1-\hat{o}_{ij}\right)^{\tau}\log(\hat{o}_{ij}) - (1-o_{ij})\hat{o}_{ij}^{\tau}\log\left(1-\hat{o}_{ij}\right), \tag{8}$$

where $\hat{o}$ is the reconstruction obtained from the decoder, hyperparameter $\tau$ controls sensitivity to the foreground, and the subscript indicates the position of image pixels.

The shared encoder is updated by the gradient from the decoder and soft Q-function, while the gradient from policy is prevented to update the encoder to improve stability [33].

### *2.3. Distributed actor and learner*

Using digital simulation, sample collection can be performed quickly and in parallel with negligible time. In such cases, RL algorithm are often applied with the paradigm that alternately collects samples and updates parameters. But in the environment which uses a real robot as shown in Section 2.1, each step of sample collection has to wait for the action execution. Therefore, both sample collection and parameter update are time-consuming. To shorten the training time, sample collection and parameter update of DSAC-AE are decoupled and performed parallelly in two processes.

The distributed structure is shown in Fig. 4. The structure is composed of an actor, a dataset, and a learner. The actor interacts with the environment and transmits transitions $(o_t, a_t, r_t, o_{t+1})$ to the dataset. The actor copies parameters from the leaner at set intervals. It is worth noting that only the encoder and

***Table I*** *Hyperparameters of DSAC-AE.*

| **Hyperparameter** | **Value** |
|---|---|
| Batch size | 512 |
| Replay buffer size | 1e4 |
| Optimizer | Adam |
| Learning rate of critic | 1e-4 |
| Learning rate of actor | 3e-5 |
| Learning rate of encoder | 1e-3 |
| Learning rate of decoder | 1e-3 |
| Hidden units (MLP) | 512 |
| Number of layers (MLP) | 2 |
| Channels (encoder) | (2,16,32,64,64) |
| Channels (decoder) | (64,64,32,16,2) |
| Discount factor $\gamma$ | 0.99 |
| Initial temperature $\alpha$ | 1.0 |
| Exponentially moving average | 3e-5 |
| Target entropy $\bar{\mathcal{H}}$ | $0.98^{*}\ln(\lvert\mathcal{A}\rvert)$ |

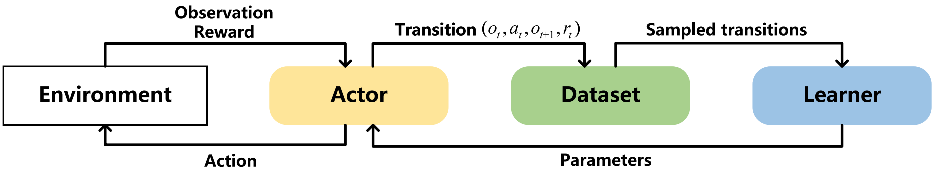

***Figure 4.*** *Distributed actor, dataset, and learner: collecting transitions, storing and sampling transitions, and updating parameters of DSAC-AE are performed distributedly by actor, dataset, and learner.*

the policy are used in the actor. The dataset acts as a common replay buffer in off-policy RL algorithms, storing transitions obtained from the actor and providing randomly sampled transitions to the learner. The learner updates parameters of the whole model as described in Section 2.2. Sample collection and parameter update are executed asynchronously. The data transmission among the actor, the dataset, and the learner is implemented using the ray package [40].

## 3. Results

Hyperparameters are set as in Table I. The vascular model is shown in Fig. 1(b), and the target is set as in Fig. 1(c). A Terumo RF*GA35153M guidewire is used in experiments. For details of the robot, we refer readers to ref. [35]. All experiments were performed with three random seeds. The robot randomly chooses actions during the first 2k sampling steps. Random actions at the beginning of training not only extensively explore the action space to avoid falling into local optimum, but also collect diverse images that are beneficial to the learning of the auto-encoder. After every 50 sampling steps, the actor copies the parameters from the learner. The learner uses a single NVIDIA TITAN Xp GPU for parameter update. The actor uses INTEL i7-10700 CPU to compute neural network without GPU acceleration.

### *3.1. Performance of DSAC-AE*

The mean reward of DSAC-AE and the baseline SAC-Discrete is shown in Fig. 5. The mean reward of DSAC-AE quickly rises to around 400 within 20k sampling steps and is about 470 after 50k sampling steps, while the mean reward of SAC-Discrete shows no upward trend and is less than 100 after 50k

https://doi.org/10.1017/S0263574722001527 Published online by Cambridge University Press

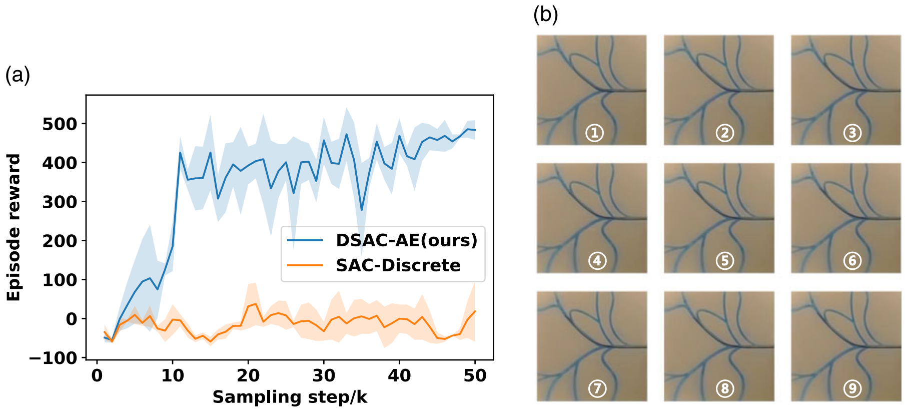

***Figure 5.*** *Performance of DSAC-AE. (a) The mean reward with DSAC-AE and SAC-Discrete. The mean reward is calculated from the last 10 trajectories during the sample collection process of actor. Light-colored parts indicate standard deviation. (b) The guidewire is oriented and passed through the bifurcation during the test.*

sampling steps. Under the reward designed in Section 2.1, the episode reward is about 530 when successfully reaching the target. Otherwise, the episode reward is less than 130. This means after 50k sampling steps, DSAC-AE could automate guidewire delivery with more than 85% probability. The results prove that the auxiliary reconstruction task in DSAC-AE can significantly improve sample efficiency.

Here is a test video[1] of DSAC-AE within 50k sampling steps. During the test, the robot directly selects the action with the highest probability. Figure 5(b) shows an example of the guidewire passing through a bifurcation in the test. The guidewire is directed towards the wrong branch at first (①). The robot can adjust the guidewire to the correct branch (②–⑧) and then deliver the guidewire through the bifurcation (⑨).

Training time and inference speed are important factors for potential clinical applications. DSAC-AE takes an average of 14.79 h (16.03, 14.08, and 14.27 h, respectively, in three experiments) for 50k sampling steps. Since all neural networks in DSAC-AE are shallow and simple, our algorithm is able to complete an action selection in about 1 ms without GPU acceleration.

### 3.2. Results with different auxiliary loss functions

Experiments with different auxiliary loss functions are designed to demonstrate the effectiveness of focal loss. Figure 6(a) shows the auxiliary loss during training. Focal loss is reduced to less than 0.001 within 1k training steps, while cross-entropy is reduced to less than 0.003 within 3k training steps. Besides, cross-entropy has larger standard deviation especially in the early 2k training steps when the robot chooses action randomly. The result indicates that focal loss converges faster and more stably than cross-entropy.

Since reconstructions cannot be quantitatively compared by values of different loss functions, several guidewire reconstruction examples after 10k training steps are shown in Fig. 6(b) for qualitative comparison. In all cases, the decoder is able to reconstruct the approximate shape of the guidewire with both auxiliary loss functions. In most cases (e.g., the first input), the decoder is able to accurately reconstruct the position and orientation of the distal tip. Since the position and orientation of the distal tip are the most critical information in guidewire manipulation, small deviation in the width of the guidewire will not affect the operation of the robot. And in the cases where the guidewire is close to borders (e.g., the second input), there is identifiable ambiguity in the distal tip. In the above cases, reconstructions with

[1] https://github.com/CASIAHaoLi/DSAC-AE-Image-Based-Reinforcement-Learning-for-Vascular-Robotic-System/blob/main/test.mp4

https://doi.org/10.1017/S0263574722001527 Published online by Cambridge University Press

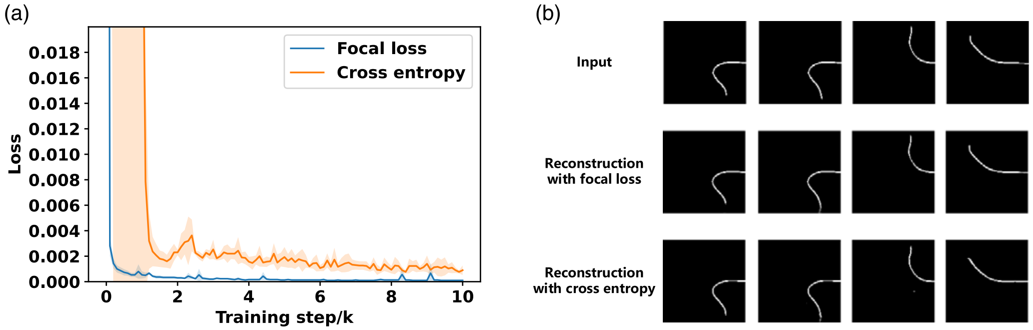

*Figure 6. Performance in the auxiliary reconstruction task. (a) The auxiliary loss within 10k training steps. (b) Several guidewire reconstruction examples after 10k training steps.*

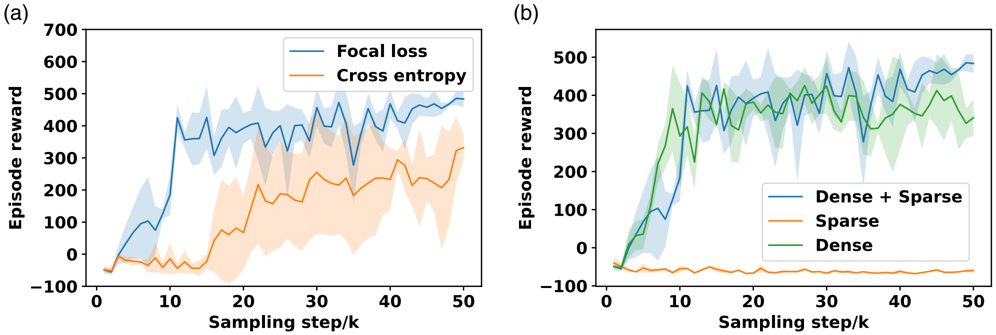

*Figure 7. The mean reward curves with different auxiliary loss functions and different sub-rewards. (a) The mean reward with different auxiliary loss function. (b) The mean reward when using different sub-rewards. Reward curves for all settings are calculated with the whole reward function for comparison.*

focal loss and those with cross-entropy have similar quality. However, details of reconstructions with cross-entropy are unsatisfactory in some cases. There may be noise spots at the background (e.g., the third input) or direction deviation of the distal tip (e.g., the fourth input).

Figure 7(a) shows the mean reward of DSAC-AE with different auxiliary loss functions. When using cross-entropy, the mean reward exceeds 300 after 50k sampling steps, indicating that DSAC-AE with cross-entropy can learn guidewire-manipulation skills but significantly slower than that with focal loss. The results demonstrate that the advantages of focal loss in the auxiliary reconstruction task can significantly improve the performance of downstream RL.

In conclusion, both DSAC-AE with focal loss and DSAC-AE with cross-entropy obtain better performance than baseline SAC-Discrete. But focal loss brings faster convergence, better stability, and better performance to both the auxiliary reconstruction task and the downstream RL.

### 3.3. Results with different reward design

The reward function guides the learning of the robot and has a great influence on the performance of RL algorithms. Ablation experiments of sparse and dense sub-rewards are designed to demonstrate their effect, while the safe sub-reward works in all settings for safety. The results are shown in Fig. 7(b). With the sparse sub-reward, the mean reward keeps lower than 0 within 50k sampling steps. With the dense sub-reward and with both the dense sub-reward and the sparse sub-reward, mean rewards are larger than 300 within 15k sampling steps, which indicates the robot is able to deliver the guidewire to the target. Under the two settings, the robot has similar performances within 30k sampling steps. However, after

https://doi.org/10.1017/S0263574722001527 Published online by Cambridge University Press

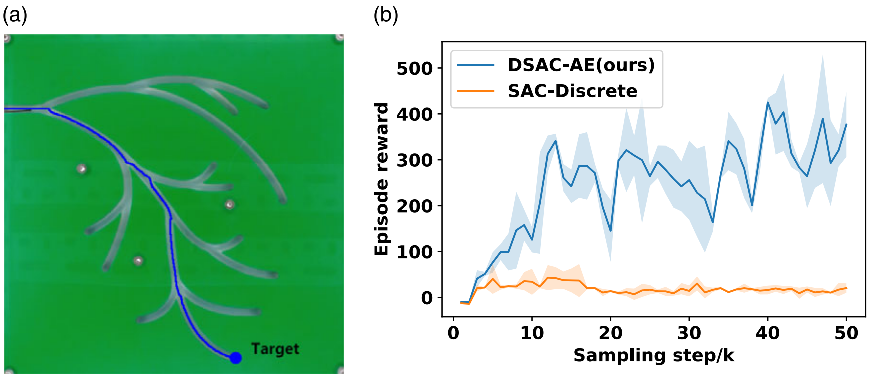

***Figure 8.*** *DSAC-AE on another vascular model. (a) The new vascular model and the target. (b) Mean reward on the vascular model.*

30k sampling steps, the mean reward with only the dense sub-reward is less than 400 and worse than that with both the dense sub-reward and the sparse sub-reward. This result shows that combining sparse sub-reward and dense sub-reward as in Section 2.1 can help the robot learn more efficiently.

### *3.4. Validation on another vascular model*

To verify the generality of our method, DSAC-AE is applied on another vascular model. The vascular model and the target are shown in Fig. 8(a). In this vascular model, vessels are more tortuous and have more complex branches, making guidewire delivery more difficult. Other settings are the same as those in Section 3.1. The mean reward of DSAC-AE and SAC-Discrete is shown in Fig. 8(b). On this complex vascular model, the mean reward of DSAC-AE is about 380 after 50k sampling steps, while the mean reward of SAC-Discrete is less than 20 after 50k sampling steps. Although the learning speed becomes slower as the task becomes more difficult, DSAC-AE still significantly outperforms SAC-Discrete. This result demonstrates the generality of DSAC-AE on different vascular models.

On this vascular model, the three experiments take an average of 12.39 h (12.38, 12.35, and 12.44 h, respectively), which is less than the mean time on the origin vascular model. This is because in the this complex vascular model, the robot cannot complete the task quickly and the environment resets less often.

## 4. Discussion

### *4.1. Analysis of temporal performance*

Training time and inference speed (i.e., speed of choosing an action) are key factors for potential clinical applications. DSAC-AE takes an average of 14.79 h for training with 50k sampling steps. Moreover, while applying the proposed algorithm in clinic, it is necessary to train the algorithm with a preoperative vascular model before performing VIS. To the best of our knowledge, there is no specific standard or statistics about the interval between obtaining preoperative vascular models and performing VIS. The interval can be as long as several days for non-acute patients, but only several hours for acute patients. Therefore, our algorithm meets training time requirements for non-acute patients, while more sample-efficient algorithms with shorter training time are needed for urgent situations. As for inference speed, DSAC-AE can choose an action based on input images in about 1 ms. The time of choosing an action is significantly shorter than the average human reaction time which is about hundreds of milliseconds [41]. Furthermore, X-ray fluoroscopy for intraoperative images is usually applied with less than 10 frames per second in clinic [42]. Consequently, the proposed algorithm meets clinical inference speed requirements.

https://doi.org/10.1017/S0263574722001527 Published online by Cambridge University Press

### 4.2. Comparison of different auxiliary loss functions

Changing input images and foreground—background imbalance are two challenges for the auxiliary reconstruction task, resulting in the advantage of focal loss. Unlike supervised learning with given data, input images in RL are collected by the robot and are constantly changing. As cross-entropy fluctuates more than focal loss after convergence, cross-entropy cannot cope with data changes well. Moreover, the change of input images is most dramatic at the beginning of training when the robot chooses action randomly for exploration. Accordingly, the variance of the cross-entropy is particularly large at the first 2k training steps. On the other hand, the foreground-background imbalance leads to the gap between the reconstruction performance of focal loss and cross-entropy. Due to foreground-background imbalance, details of the vessel and the guidewire are hard to predict. Furthermore, the pixels hard to predict have higher weights in focal loss as shown in Eq. (8), while in cross-entropy all pixels have the same weight. Therefore, focal loss pays more attention to details and has better reconstruction performance with foreground-background imbalance.

In the cases where the guidewire is close to borders, reconstructions with focal loss and those with cross-entropy both have identifiable ambiguity in the distal tip. This may be because, in addition to the auxiliary reconstruction task, soft Bellman residual (i.e., Eq. (1)) is also used to update the encoder. Since trajectories end when the distal tip reaches borders and the dense reward does not work when the distal tip is in the wrong branch, those inputs where the distal tip is close to borders have similar Q-functions. Thus there may be similarities in the embedding of these inputs, which cause the ambiguity in reconstructions.

### 4.3. Impact of different sub-rewards

With sparse sub-reward or dense sub-reward alone, the robot performs worse than with both sub-rewards. The sparse reward is zero and has no useful information unless the target is reached. Moreover, the robot is in near-random exploration until obtaining useful information. Thus, when only using sparse sub-reward, the robot must first reach the target with a near-random exploration to obtain useful samples with no-zero rewards. This is almost impossible in our scenario. Consequently, the robot learns almost nothing with sparse sub-reward alone. With the dense sub-reward alone and with both dense and sparse sub-rewards, the robot has similar performance in 30k sampling steps. This is because the sparse sub-reward is always zero and has little effect until learning a passable strategy that can often reach the target. After learning a passable strategy, the sparse sub-reward makes actions that are effective to reach the target have larger Q-functions, reinforcing the probability of those actions. Therefore, with both dense and sparse sub-rewards, the robot has better final performance than with the dense sub-reward alone.

## 5. Conclusions and future work

A novel RL algorithm DSAC-AE is proposed to automate guidewire delivery on realistic vascular robotic systems. Using auxiliary reconstruction task, DSAC-AE has high sample efficiency with high-dimensional image input. The use of focal loss in DSAC-AE can effectively mitigate the impact of foreground-background imbalance. Besides, distributed sample collection and parameter update can further shorten training time. The task can be completed within an acceptable amount (50k) of sampling steps in less than 15 h. In the subsequent work, generalization among various instruments and diverse vascular models will be considered. In addition, we will try to combine prior data such as images and human demonstrations with our algorithm to further improve the sample efficiency and meet stricter training time requirements in clinic.

**Author contributions.** Zeng-Guang Hou and Hao Li conceived the study; Hao Li completed the code and data analysis; Hao Li, Xiao-Hu Zhou, Xiao-Liang Xie, and Shi-Qi Liu designed overall experiments; Mei-Jiang Gui, Tian-Yu Xiang, Jin-Li Wang debugged the experimental environment; Zeng-Guang Hou and Xiao-Hu Zhou reviewed the manuscript critically. All authors read and approved the manuscript.

**Financial support.** This work was supported in part by the National Natural Science Foundation of China under Grant 62003343, Grant 62222316, Grant U1913601, Grant 62073325, Grant 61720106012, Grant U20A20224, and Grant U1913210; in part by the Beijing Natural Science Foundation under Grant M22008; in part by the Youth Innovation Promotion Association of Chinese Academy of Sciences (CAS) under Grant 2020140 and in part by the Strategic Priority Research Program of CAS under Grant XDB32040000.

**Conflicts of interest.** The authors declare no conflicts of interest exist.

**Ethical Approval.** Not applicable.

## References

[1] G. A. Mensah, G. A. Roth and V. Fuster, "The global burden of cardiovascular diseases and risk factors: 2020 and beyond," *J. Am. Coll. Cardiol.* **74**(20), 2529–2532 (2019).

[2] X.-H. Zhou, X.-L. Xie, S.-Q. Liu, Z.-L. Ni, Y.-J. Zhou, R.-Q. Li, M.-J. Gui, C.-C. Fan, Z.-Q. Feng, G.-B. Bian and Z.-G. Hou, "Learning skill characteristics from manipulations," *IEEE Trans. Neural Netw. Learn. Syst.*, 1–15 (2022). doi: 10.1109/TNNLS.2022.3160159.

[3] H. Rafii-Tari, C. J. Payne and G.-Z. Yang, "Current and emerging robot-assisted endovascular catheterization technologies: A review," *Ann. Biomed. Eng.* **42**(4), 697–715 (2014).

[4] X.-H. Zhou, X.-L. Xie, S.-Q. Liu, Z.-Q. Feng, M.-J. Gui, J.-L. Wang, H. Li, T.-Y. Xiang, G.-B. Bian and Z.-G. Hou, "Surgical skill assessment based on dynamic warping manipulations," *IEEE Trans. Med. Robot. Bionics.* **4**(1), 50–61 (2022).

[5] A. Roguin, J. Goldstein and O. Bar, "Brain tumours among interventional cardiologists: A cause for alarm? Report of four new cases from two cities and a review of the literature," *EuroIntervention* **7**(9), 1081–1086 (2012).

[6] A. Karatasakis, H. S. Brilakis, B. A. Danek, J. Karacsonyi, J. R. Martinez-Parachini, P. J. Nguyen-Trong, A. J. Alame, M. K. Roesle, B. V. Rangan, K. Rosenfield, R. Mehran, E. Mahmud, C. E. Chambers, S. Banerjee and E. S. Brilakis, "Radiation-associated lens changes in the cardiac catheterization laboratory: Results from the IC-CATARACT(CATaracts Attributed to RAdiation in the CaTh lab) study," *Catheter. Cardiovasc. Interv.* **91**(4), 647–654 (2018).

[7] A. Elmaraezy, M. E. Morra, A. T. Mohammed, A. AlHabaa, A. S. Elgebaly, A. A. Ghazy, A. Khalil, N. T. Huy and K. Hirayama, "Risk of cataract among interventional cardiologists and catheterization lab staff: A systematic review and meta-analysis," *Catheter. Cardiovasc. Interv.* **90**(1), 1–9 (2017).

[8] L. W. Klein, Y. Tra, K. N. Garratt, W. A. Powell, G. Lopez-Cruz, C. E. Chambers and J. A. Goldstein, "Occupational health hazards of interventional cardiologists in the current decade: Results of the 2014 SCAI membership survey," *Catheter. Cardiovasc. Interv.* **86**(5), 913–924 (2015).

[9] J. F. Granada, J. A. Delgado, M. P. Uribe, A. Fernández, G. Blanco, M. B. Leon and G. Weisz, "First-in-human evaluation of a novel robotic-assisted coronary angioplasty system," *J. Am. Coll. Cardiol. Cardiovas. Interv.* **4**(4), 460–465 (2011).

[10] J.-H. Woo, H.-S. Song, H.-J. Cha and B.-J. Yi, "Advantage of steerable catheter and haptic feedback for a 5-DOF vascular intervention robot system," *Appl. Sci.* **9**(20), 4305 (2019).

[11] S. Guo, Y. Song, X. Yin, L. Zhang, T. Tamiya, H. Hirata and H. Ishihara, "A novel robot-assisted endovascular catheterization system with haptic force feedback," *IEEE Trans. Robot.* **35**, 685–696 (2019).

[12] M.-J. Gui, X.-H. Zhou, X.-L. Xie, S.-Q. Liu, L. H., T.-Y. Xia, J.-L. Wang and Z.-G. Hou, "Design and experiments of a novel Halbach-cylinder-based magnetic skin: A preliminary study," *IEEE Trans. Instrum. Meas.* **71**, 1–11 (2022). doi: 10.1109/TIM.2022.3147904.

[13] T. M. Patel, S. Shah and S. B. Pancholy, "Long distance tele-robotic-assisted percutaneous coronary intervention: A report of first-in-human experience," *EClinicalMedicine* **14**, 53–58 (2019).

[14] G.-Z. Yang, J. Cambias, K. Cleary, E. Daimler, J. Drake, P. E. Dupont, N. Hata, P. Kazanzides, S. Martel, R. V. Patel, V. J. Santos and R. H. Taylor, "Medical robotics-regulatory, ethical, and legal considerations for increasing levels of autonomy," *Sci. Robot.* **2**(4), eaam8638 (2017).

[15] A. A. Nooryani and W. Aboushokka, "Rotate-on-retract procedural automation for robotic-assisted percutaneous coronary intervention: First clinical experience," *Case Rep. Cardiol.* **2018**, 1–3 (2018).

[16] H. Rafii-Tari, J. Liu, S.-L. Lee, C. D. Bicknell and G.-Z. Yang, "Learning-Based Modeling of Endovascular Navigation for Collaborative Robotic Catheterization," **In:** *Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention* (Springer, Berlin, 2013) pp. 369–377.

[17] H. Rafii-Tari, J. Liu, C. J. Payne, C. D. Bicknell and G.-Z. Yang, "Hierarchical HMM Based Learning of Navigation Primitives for Cooperative Robotic Endovascular Catheterization," **In:** *Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention* (Springer, Berlin, 2014) pp. 496–503.

[18] W. Chi, J. Liu, H. Rafii-Tari, C. V. Riga, C. D. Bicknell and G.- Z. Yang, "Learning-based endovascular navigation through the use of non-rigid registration for collaborative robotic catheterization," *Int. J. Comput. Assist. Radiol. Surg.* **13**(6), 855–864 (2018).

[19] Y. Zhao, S. Guo, Y. Wang, J. Cui, Y. Ma, Y. Zeng, X. Liu, Y. Jiang, Y. Li, L. Shi and N. Xiao, "A CNN-based prototype method of unstructured surgical state perception and navigation for an endovascular surgery robot," *Med. Biol. Eng. Comput.* **57**(9), 1875–1887 (2019).

[20] J. Guo, S. Feng and S. Guo, "Study on the Automatic Surgical Method of the Vascular Interventional Surgical Robot Based on Deep Learning," **In:** *Proceedings of 2021 IEEE International Conference on Mechatronics and Automation* (Institute of Electrical and Electronics Engineers Inc., Piscataway, 2021) pp. 1076–1081.

https://doi.org/10.1017/S0263574722001527 Published online by Cambridge University Press

[21] W. Chi, G. Dagnino, T. M. Y. Kwok, A. Nguyen, D. Kundrat, M. E. M. K. Abdelaziz, C. Riga, C. Bicknell and G.-Z. Yang, "Collaborative Robot-Assisted Endovascular Catheterization with Generative Adversarial Imitation Learning," **In:** *Proceedings of 2020 IEEE International Conference on Robotics and Automation* (Institute of Electrical and Electronics Engineers Inc., Piscataway, 2020) pp. 2414–2420.

[22] J. Schrittwieser, I. Antonoglou, T. Hubert, K. Simonyan, L. Sifre, S. Schmitt, A. Guez, E. Lockhart, D. Hassabis, T. Graepel, T. Lillicrap and D. Silver, "Mastering Atari, Go, chess and shogi by planning with a learned model," *Nature* **588**(7839), 604–609 (2020).

[23] V. Azimirad and M. F. Sani, "Experimental study of reinforcement learning in mobile robots through spiking architecture of Thalamo-Cortico-Thalamic circuitry of mammalian brain," *Robotica* **38**(9), 1558–1575 (2020).

[24] M. Gómez, R. V. González, T. Martínez-Marín, D. Meziat and S. Sánchez, "Optimal motion planning by reinforcement learning in autonomous mobile vehicles," *Robotica* **30**(2), 159–170 (2012).

[25] L. Karstensen, T. Behr, T. P. Pusch, F. Mathis-Ullrich and J. Stallkamp, "Autonomous guidewire navigation in a two dimensional vascular phantom," *Curr. Dir. Biomed. Eng.* **6**(1), 20200007 (2020).

[26] T. Behr, T. P. Pusch, M. Siegfarth, D. Hüsener, T. Mörschel and L. Karstensen, "Deep reinforcement learning for the navigation of neurovascular catheters," *Curr. Dir. Biomed. Eng.* **5**(1), 5–8 (2019).

[27] W. Chi, J. Liu, M. E. M. K. Abdelaziz, G. Dagnino, C. V. Riga, C. D. Bicknell and G.-Z. Yang, "Trajectory Optimization of Robot-Assisted Endovascular Catheterization with Reinforcement Learning," **In:** *Proceedings of 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems* (Institute of Electrical and Electronics Engineers Inc., Piscataway, 2018) pp. 3875–3881.

[28] C. Shi, X. Luo, J. Guo, Z. Najdovski, T. Fukuda and H. Ren, "Three-dimensional intravascular reconstruction techniques based on intravascular ultrasound: A technical review," *IEEE J. Biomed. Health Inform.* **22**(3), 806–817 (2018).

[29] Z. Wang, N. de Freitas and M. Lanctot, "Dueling network architectures for deep reinforcement learning," *CoRR*, abs/1511.06581 (2015).

[30] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. P. Lillicrap, T. Harley, D. Silver and K. Kavukcuoglu, "Asynchronous Methods for Deep Reinforcement Learning," **In:** *Proceedings of the 33rd International Conference on Machine Learning* (Proceedings of Machine Learning Research, New York, 2016) pp. 1928–1937.

[31] H. You, E. Bae, Y. Moon, J. Kweon and J. Choi, "Automatic control of cardiac ablation catheter with deep reinforcement learning method," *J. Mech. Sci. Technol.* **33**(11), 5415–5423 (2019).

[32] F. Meng, S. Guo, W. Zhou and Z. Chen, "Evaluation of a Reinforcement Learning Algorithm for Vascular Intervention Surgery," **In:** *Proceedings of 2021 IEEE International Conference on Mechatronics and Automation* (Institute of Electrical and Electronics Engineers Inc., Piscataway, 2021) pp. 1033–1037.

[33] D. Yarats, A. Zhang, I. Kostrikov, B. Amos, J. Pineau and R. Fergus, "Improving Sample Efficiency in Model-Free Reinforcement Learning from Images," **In:** *Proceedings of the 35th AAAI Conference on Artificial Intelligence* (Association for the Advancement of Artificial Intelligence, Menlo Park, 2021) pp. 10674–10681.

[34] A. Srinivas, M. Laskin and P. Abbeel, "CURL: Contrastive Unsupervised Representations for Reinforcement Learning," **In:** *Proceedings of the 37th International Conference on Machine Learning* (Proceedings of Machine Learning Research, NewYork, 2020) pp. 5639–5650.

[35] H.-L. Zhao, S.-Q. Liu, X.-H. Zhou, X.-L. Xie, Z.-G. Hou, Y.-J. Zhou, L.-S. Zhang, M.-J. Gui and J.-L. Wang, "Design and Performance Evaluation of a Novel Vascular Robotic System for Complex Percutaneous Coronary Interventions," **In:** *Proceedings of 2021 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (Institute of Electrical and Electronics Engineers Inc., Piscataway, 2021) pp. 4679–4682.

[36] B. D. Ziebart, A. L. Maas, J. A. Bagnell and A. K. Dey, "Maximum Entropy Inverse Reinforcement Learning," **In:** *Proceedings of the 23rd AAAI Conference on Artificial Intelligence* (Association for the Advancement of Artificial Intelligence, Menlo Park, 2008) pp. 1433–1438.

[37] T. Haarnoja, A. Zhou, K. Hartikainen, G. Tucker, S. Ha, J. Tan, V. Kumar, H. Zhu, A. Gupta, P. Abbeel and S. Levine, "Soft actor-critic algorithms and applications," *CoRR*, abs/1812.05905 (2018).

[38] P. Christodoulou, "Soft actor-critic for discrete action settings," *CoRR*, abs/1910.07207 (2019).

[39] T.-Y. Lin, P. Goyal, R. B. Girshick, K. He and P. Dollár, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.* **42**, 318–327 (2020).

[40] P. Moritz, R. Nishihara, S. Wang, A. Tumanov, R. Liaw, E. Liang, W. Paul, M. I. Jordan and I. Stoica, "Ray: A distributed framework for emerging ai applications," *CoRR*, abs/1712.05889 (2017).

[41] E. Lew, S. S. Ricardo Chavarriaga and J. Millán, "Detection of self-paced reaching movement intention from EEG signals," *Front. Neuroeng.* **5**, 13 (2012).

[42] H. Heidbuchel, F. H. M. Wittkampf, E. Vañó, S. Ernst, R. J. Schilling, E. Picano, L. Mont, P. Jais, J. de Bono, C. Piorkowski, E. B. Saad and F. J. Femenía, "Practical ways to reduce radiation dose for patients and staff during device implantations and electrophysiological procedures," *Europace* **16**(7), 946–964 (2014).

**Cite this article:** H. Li, X.-H. Zhou, X.-L. Xie, S.-Q. Liu, M.-J. Gui, T.-Y. Xiang, J.-L. Wang and Z.-G. Hou (2023). "Discrete soft actor-critic with auto-encoder on vascular robotic system", Robotica **41**, 1115–1126. https://doi.org/10.1017/S0263574722001527