Journal of Ambient Intelligence and Humanized Computing (2022) 13:5011–5022
https://doi.org/10.1007/s12652-020-01905-3

ORIGINAL RESEARCH

Multi-step medical image segmentation based on reinforcement learning

Zhiqiang Tian[1] · Xiangyu Si[1] · Yaoyue Zheng[1] · Zhang Chen[1] · Xiaojian Li[1]

Received: 22 December 2019 / Accepted: 19 March 2020 / Published online: 27 March 2020

Abstract
Image segmentation technology has made a remarkable effect in medical image analysis and processing, which is used to help physicians get a more accurate diagnosis. Manual segmentation of the medical image requires a lot of effort by professionals, which is also a subjective task. Therefore, developing an advanced segmentation method is an essential demand. We propose an end-to-end segmentation method for medical images, which mimics physicians delineating a region of interest (ROI) on the medical image in a multi-step manner. This multi-step operation improves the performance from a coarse result to a fine result progressively. In this paper, the segmentation process is formulated as a Markov decision process and solved by a deep reinforcement learning (DRL) algorithm, which trains an agent for segmenting ROI in images. The agent performs a serial action to delineate the ROI. We define the action as a set of continuous parameters. Then, we adopted a DRL algorithm called deep deterministic policy gradient to learn the segmentation model in continuous action space. The experimental result shows that the proposed method has 7.24% improved to the state-of-the-art method on three prostate MR data sets and has 3.52% improved on one retinal fundus image data set.

Keywords Reinforcement learning · Deep deterministic policy gradient · Image segmentation · Multi-step manner

1 Introduction

Segmentation of medical image plays an important role in diagnosis and treatment. Based on the segmentation of region of interest, physicians can diagnose potential disease, understand the nature of the lesion, get the location and scope of the attack. Image segmentation is a widely used image processing technique, which aims to divide an image into two or more meaning regions (Long et al. 2015; Eltanboly et al. 2019; Aguirreramos et al. 2018; Ahmadvand and Daliri 2015; Wu et al. 2019). Recently, a rapid development of deep learning-based image segmentation occurs in biological fields (Litjens et al. 2017; Ronneberger et al. 2015). These methods got more and more accurate results by designing various architectures of deep neural networks, feature encoders, and exploring contextual relationships between objects and background. Long et al. (2015) firstly proposed a fully convolutional network (FCN) to directly segment a whole image based on traditional classification network by using a skip architecture. Since then, many automatic segmentation methods were proposed based on FCN. Ronneberger et al. (2015) designed a more elaborate network for medical images. Instead of the skip architecture, they used a more symmetrical encoder-decoder architecture and un-sample the feature to original size layer by layer, which makes the segmentation result more accurate. Milletari et al. (2016) designed a V-Net, which extends the 2D segmentation task to 3D medical images. To get more accurate segmentation results, some interactive segmentation methods were adopted with the interaction of physicians. Grabcut (Rother et al. 2004) is a classical interactive segmentation method, which needs a coarse bounding box to distinguish foreground and background. Guotai et al. (2018) use several scribbles on initial segmentation result of CNN model to further fine-tune the result.

We observed that when physicians delineate the ROI on a medical image, a coarse segmentation is performed first, which will identify most of the ROI area. Then, the coarse segmentation is refined in a multi-step manner by physicians. Such a multi-step segmentation process is similar to interactive segmentation (Rother et al. 2004). The

✉ Zhiqiang Tian
zhiqiangtian@xjtu.edu.cn

[1] School of Software Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi, China

interaction can be considered as a prior knowledge for segmentation, which includes drawing several strokes on the foreground and background with a brush or drawing a box around the foreground. This technique can gain prior knowledge by interacting with user to improve the performance of segmentation algorithm. Inspired by these concepts, the proposed method presents a multi-step segmentation algorithm. The segmentation of every step is based on the previous segmentation mask. The advantage of our method is that it can automatically obtain prior knowledge to further improve the segmentation performance without interaction.

In recent years, Reinforcement Learning (RL) has been widely used in various artificial intelligence issues, including computer vision, robot control, anomaly detection, automatic driving, and computer game. With the breakthrough of deep learning (DL), the researchers combine it with RL for solving more complex problem. The combination of DL and RL forms a deep reinforcement learning (DRL) algorithm, which is used to train an intelligent agent to solve the MDP problem. In this paper, deep reinforcement learning (Arulkumaran et al. 2017) is adopted to implement the multi-step medical image segmentation.

We propose an automatic multi-step method based on DRL for medical image segmentation. The segmentation process is formulated as a Markov decision process and solved by a DRL algorithm. During the segmentation process, the previous predicted segmentation mask is used as a prior knowledge for the next step. For each step, the agent performs a segmentation action based on the input image and current segmentation mask. Inspired by stroke-based stylization method (Xie et al. 2015), we present a segmentation executor to draw a brushstroke on the input segmentation mask, which indicates the ROI. The segmentation executor is an end-to-end neural network, which maps action parameters to the brushstroke. The executor can be implemented in various shapes. After pre-set number of steps, we can get the final segmentation result.

We define a set of continuous action parameters to control the location and shape of a brushstroke for fine-grained segmentation. This segmentation process is formulated as a sequential decision-making problem and optimize it with DRL. Deep Q-Network (DQN) (Mnih et al. 2015) is one of the most widely used DRL algorithms. However, DQN can only solve MDP in discrete action space. To address this problem, we adopt DDPG algorithm to solve it in a continuous action space.

To the best of our knowledge, this is the first study to formulate medical image segmentation problem as an MDP, and solve it by DDPG. The contributions of this paper are summarized as follows. (1) We define medical image segmentation as a Markov decision process and solve it by deep deterministic policy gradient, which mimics the physician delineating the ROIs on medical images. (2) We present a quadratic Bezier curve (QBC) based segmentation executor for medical image segmentation, and use an action bundle strategy to further improve the segmentation accuracy. (3) To better train a segmentation agent, we propose a modified experience replay memory (ERM) for robust segmentation.

The following of the paper is organized as follows. Section 2 overviews related works. The proposed method is introduced in Sect. 3. In Sect. 4, the segmentation results are presented. The conclusions are given in Sect. 5.

2 Related work

Most of the current image segmentation methods are based on fully convolutional neural network (FCN) (Long et al. 2015), which uses pixel-level classification technique for image segmentation. Inspired by FCN, encoder-decoder architecture is widely used for image segmentation, such as U-Net (Ronneberger et al. 2015), V-Net (Milletari et al. 2016), and DeepLab (Chen et al. 2017). Encoder is typically used to extract feature and reduce spatial dimension, while decoder is typically used to progressively restore target and spatial dimension information and directly outputs the final segmentation mask. The architecture of U-Net is similar to the FCN, which is divided into the sub-sampling stage and the up-sampling stage. The U-net uses the skip connection structure to connect the lower layer to the upper layer. Therefore, the features extracted by the lower layer can be passed directly to the upper layer, which makes the pixel positioning of the U-net network more accurate. For 3D image segmentation, a three-dimensional convolution is used in V-Net. A Dice coefficient based loss function is proposed to optimize the model. Besides the FCN-based methods, many deep learning-based segmentation methods are also proposed for image segmentation, such as Polygon-RNN (Castrejon et al. 2017), DeepLab V3+ (Chen et al. 2018), and Multi-task Network Cascades (Dai et al. 2016). In recent years, many new algorithms have also been proposed for various tasks, such as regions extraction (Lu et al. 2020), wound intensity correction (Lu et al. 2017), and automatic classification of lung nodules (Yoshino et al. 2017).

Although above-mentioned methods could get satisfactory results, only few works explore the process of physicians delineating the region of interest on medical images. The RL can be used to mimic the delineation process of physicians. In recent years, reinforcement learning has made remarkable achievements in a wide range of applications by combining it with deep learning. DRL methods use deep neural networks for agent training, such as Deep Q-Network, Deep Deterministic Policy Gradient (DDPG) (Lillicrap et al. 2015), Proximal Policy Optimization (PPO) (Schulman et al. 2017), and Asynchronous Advantage Actor-Critic(A3C) (Mnih et al. 2016). DeepMind achieves human-level player skill (Silver et al. 2016) in playing games by DRL. Therefore, more researchers began to apply DRL in many problems, such as recommendation system (Jagadeesan and Subbiah 2020; Madani et al. 2019), game simulator (Zhu and Zhao 2019), and internet of things (Kim et al. 2019). In addition, DRL has shown great potential in many challenging tasks such as image classification problems (Ba et al. 2014), landmark detection (Alansary et al. 2019), object localization (Caicedo and Lazebnik 2015), visual navigation (Zhu et al. 2017), semantic parsing of large scale 3D point cloud (Liu et al. 2017), and face recognition (Rao et al. 2017).

In this paper, we propose a segmentation method based on DRL for medical image segmentation. The key of our method is how to formulate the problem as an MDP. Sahba et al. (2008) proposes a reinforcement learning method for prostate image segmentation. Q-learning (Watkins and Dayan 1992) is used to find an appropriate local values for sub-images and to extract the prostate from the image. However, Q-learning only can handle small space of state and action. In recent years, some researchers tried to use Deep Q-Network, which combines Q-learning with convolutional neural networks (CNN) for image segmentation. DeepOutline (Wang et al. 2018) is an end-to-end deep reinforcement learning network for semantic image segmentation, which copies a user holding a pen to draw the outline of objects in the image. This process is formulated as an MDP. SeedNet (Song et al. 2018) presents a novel automatic seed generation system for the task of interactive segmentation. Both of these methods use DQN to train an agent for image segmentation. However, DQN cannot handle continuous action, which needs additional efforts to deal with the problem. In this paper, we directly use DDPG for image segmentation to avoid additional efforts.

3 Method

3.1 Overview

In this work, we propose an automatic multi-step segmentation method based on deep reinforcement learning for medical images. A segmentation agent is trained in each step to get an optimized segmentation policy based on the evaluation of the current step. In this paper, DDPG algorithm is used to train the segmentation agent for solving the MDP problem. Here, deep deterministic policy gradient algorithm is a combination of deterministic policy gradient (DPG) (Silver et al. 2014) and deep learning. Compared with the traditional method, e.g., level set, snakes, chanvese, the proposed method needs no professional experience. The proposed method can be optimized by neural network according to the segmentation results of the previous step. The neural network can learn the appropriate strategy for medical image segmentation without professional experience.

DDPG is based on an Actor-Critic (AC) framework, which can be used to solve the problems of the continuous action space. The Actor-Critic framework is a combination of the policy-based and value-based methods. The former uses an indirect method to learn a value function or an action value function to get the policy. In contrast, the latter directly model the policy, such as policy gradient (PG) (Sutton et al. 2000), which is known as policy optimization. The Actor-Critic contains two networks, which are policy network and value network. The policy network is called actor, while the value network is called critic. In the AC framework, the actor is responsible for learning policy, while the critic is responsible for making evaluation for the decision of actor. The goal of actor is to get a better rating, and the goal of critic is to be more accurate. Since actor and critic are interdependent and interact with each other, an iterative optimization is adopted during training, which adopts the idea of adversarial networks (Xu et al. 2019).

Moreover, our method includes an off-the-shelf segmentation executor, which performs the segmentation action based on a set of continuous action parameters. The segmentation executor outputs a brushstroke based on a set of action parameters and draw it on the input segmentation mask to further refine the segmentation results.

The overall architecture of our method is shown in Fig. 1. The proposed method includes an segmentation executor, which is implemented by using neural networks to perform segmentation action. Given a medical image and an initial segmentation mask, the agent aims to find an action sequence $(A_0, \ldots, A_t, A_{t+1}, \ldots, A_T)$ for segmentation task based on current segmentation policy $\pi(S)$ (mapping of state S to action A). Selecting a suitable action in each step is very important, i.e. this decision should be compatible with previous and future decision. The quality of the segmentation policy directly determines the accuracy of the segmentation result. The segmentation policy can be obtained by training the agent with DDPG. In step t, the segmentation executor adopts A_t selected by actor to generate a brushstroke and renders it on the segmentation mask M_t to get an updated segmentation mask M_{t+1}. These operations are repeated throughout the segmentation process. Finally, we get the final segmentation mask with the learned segmentation policy π.

A residual structure is adopted similar to ResNet-18 (He et al. 2016) for the value network (critic) and policy network (actor). Meanwhile, batch normalization (Ioffe and Szegedy 2015) is used for the policy network. To better train the model, weight normalization (Salimans and Kingma 2016) with Translated ReLU (TReLU) (Xiang and Li 2017) is added for the value network. The segmentation network consists of fully connected layers and convolution layers.

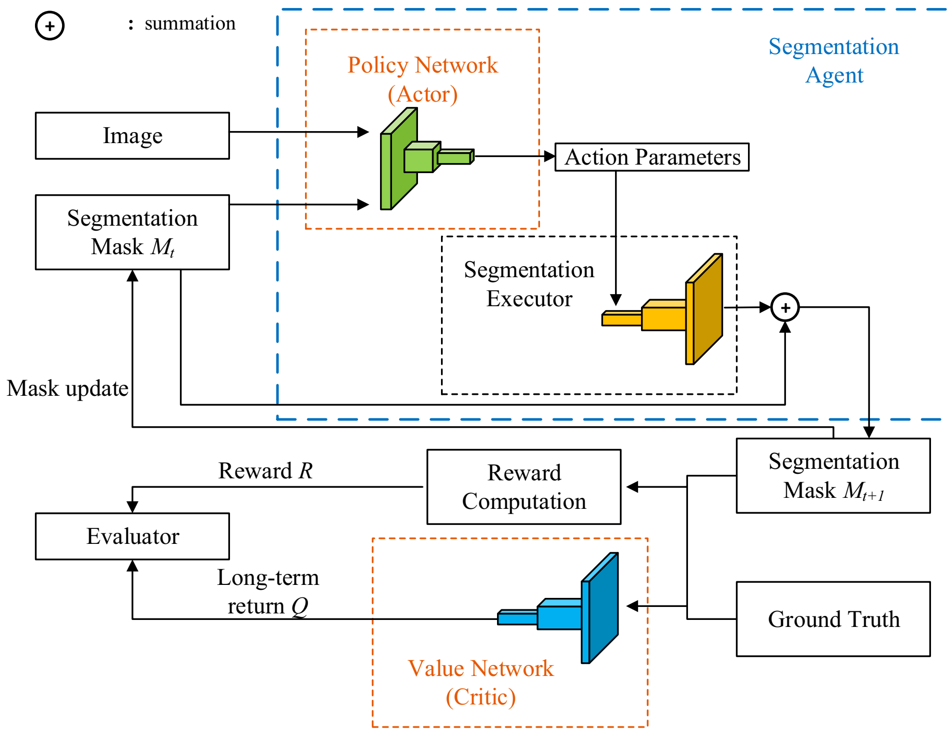


Fig. 1 Overview of the proposed method. Given a medical image and an initial segmentation mask, the actor selects a set of action parameters based on the image and current segmentation mask M_t in the step t. The parameters are fed into segmentation executor. Then, the segmentation executor generates an updated segmentation mask M_{t+1}. The updated segmentation mask has three roles. First, it is used to calculate the reward by comparing it with ground truth mask. Second, it is fed into critic with ground truth to get a long-term expected return Q. Third, it is used to update the previous segmentation mask M_t. The actor, critic, and segmentation executor are implemented by neural networks. Evaluator will further evaluate the current segmentation policy π based on the long-term expected return Q and reward R

Sub-pixel (Shi et al. 2016) strategy is used to increase the resolution of brushstrokes in the segmentation executor. In addition, CoordConv (Liu et al. 2018) is used as the first layer in actor and critic. The network structures of actor, critic, and segmentation executor are shown in Fig. 2.

We will explain how we define the image segmentation problem as an MDP process in Sect. 3.2. In Sect. 3.3, we introduce the detail of segmentation executor and a strategy called action bundle (Huang et al. 2019) adopted in our method to improve the accuracy. In Sect. 3.4, a modified experience

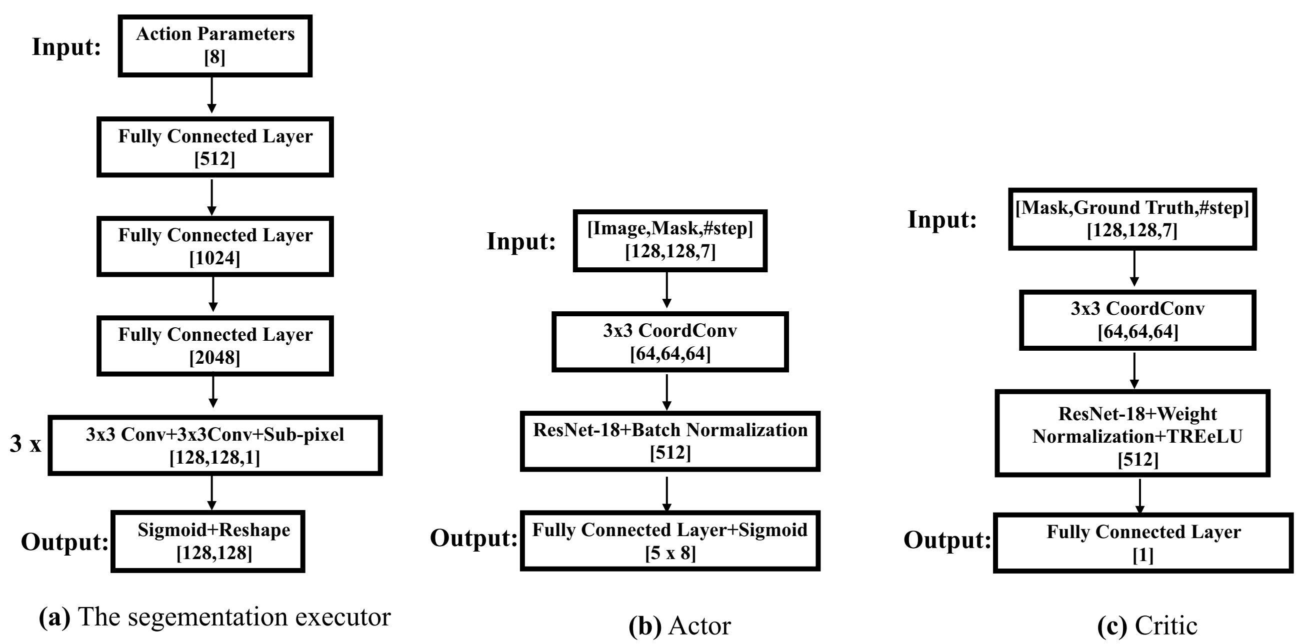


Fig. 2 Network structures. **a** The segmentation executor output a revised segmentation results based on action parameters. **b** Because the action bundle value is 5, the actor outputs five sets of action parameters at a time, each containing eight parameters. **c** The critic outputs only one evaluation value. Resnet-18 is used in the network

replay memory is proposed for DDPG training, which can further improve the accuracy.

3.2 Markov decision process for segmentation

In this paper, the medical image segmentation is casted as a Markov decision process, where a segmentation agent finds out the ROI. The MDP has three key parts, which are state space S, action A, and reward function. The definitions of these three parts in our method are presented as follows.

State: The state space contains all the information that the agent can observe in the environment. It is served as the basis for the actor to select the action. In this work, a state contains the current segmentation mask M_t, image I, and step index t, which denotes as $S_t = (M_t, I, t)$. M_t is segmentation mask that pixel value is 0 or 255. The pixels in the background area are 0, and 255 in the foreground. The default value of the initial segmentation mask is 0. I is the medical image to be segmented. The step index t is used to distinguish each step in the segmentation process. There is a terminal state in our multi-step segmentation process. A maximum number of steps needs to be set before training. When the step reaches the maximum number of steps, the agent enters the terminal state, which could get final segmentation result.

Action: The action space contains all the actions that segmentation executor can perform. Given a state, the agent selects an action in the action space based on a policy π. Then, the action is used to control the position and shape of the brushstroke, which is defined as a set of parameters.

Reward function: In reinforcement learning task, reward function is a mapping from state to reward. The task of the agent is to continuously maximize the sum of discounted future rewards R. The reward function refers to the immediate reward of a state changed after an action, which is used to evaluate the effectiveness of the result for the agent decision. In the training state, the segmentation mask M is updated in each step. Therefore, the accuracy of the mask can be obtained by comparing it with the ground truth mask in each step. We adopt a mean square error ($L2$) as the evaluation metric. The reward function with $L2$ loss is described as R_{l2},

$$R_{l2} = L2(M, G). \tag{1}$$

In order to better represent the effect of each step, we need another basic reward function using the change trend of $L2$. $L2$ value measures the similarity between two images. If two images are similar, the $L2$ loss would be close to 0. A reward function can be designed by using the variation of R_{l2} between two adjacent steps. The reward function is described as R_{diff},

$$R_{diff} = L2(M_{t-1}, G) - L2(M_t, G), \tag{2}$$

where M_{t-1} denotes the segmentation mask of the previous step and M_t denotes the current mask. The reward function gives a positive signal if the $L2$ loss is decreased, and vice versa. In reinforcement learning tasks, we also need to estimate the long-term return Q value for each step. The reward function presents the quality of the selected action in each step. The value function represents the quality of the selected action for the whole segmentation process. Once the reward is obtained, it can be used to calculate the $Q(S_t, A_t)$ based on Bellman equation,

$$Q(S_t, A_t) = R(S_t, A_t) + \gamma Q(S_{t+1}, \pi(S_{t+1})), \tag{3}$$

where $Q(S_t, A_t)$ represents the Q value of selecting A_t under state S_t. $R(S_t, A_t)$ represents the reward function. γ is a discount factor, which indicates the importance of the future returns $Q(S_{t+1}, \pi(S_{t+1}))$ compared with the immediate reward $R(S_t, A_t)$. When γ is 0, it equivalent to only considering immediate reward without considering long-term returns. When γ is 1, long-term returns and immediate reward are equally important. π is the segmentation policy. The critic estimates the long-term return Q for the agent decision, which is learned by using Bellman equation.

The original Bellman equation estimates the Q value based on the state and action. To improve the evaluation accuracy of critic, S_t and ground truth are fed into critic rather than S_t and A_t. The modified value function $V(S_t, G)$ is learned by using the following equation,

$$V(S_t, G) = R(S_t, A_t) + \gamma V(S_{t+1}, G). \tag{4}$$

Finally, we adopt the DDPG to optimize the MDP for medical image segmentation.

3.3 Segmentation executor and action bundle

Above-mentioned segmentation executor is implemented by a neural network, which draws a brushstroke on the mask as a renderer to indicate the ROI. There are two advantages of using neural networks for segmentation executor. First, it is differentiable, which can be well combined with DDPG. Second, fine-grained actions can be performed by the neural network. The segmentation executor is trained by using supervised learning on lots of training samples, which are obtained by graphical renderer programs. Several segmentation executors are used to generate different shapes of brushstroke, which includes triangle, round, and quadratic Bezier curve. Based on the experimental results, QBC can get best performance for the medical image segmentation. Therefore, in this prostate image segmentation task, we only use the Bezier curve. The action parameters of QBC are defined as follows,

$$A_t = (x_0, y_0, x_1, y_1, x_2, y_2, r_0, r_1), \tag{5}$$

where $(x_0, y_0, x_1, y_1, x_2, y_2)$ are the coordinates of three control points (P_0, P_1, P_2) of the QBC. The parameters (r_0, r_1) control the thickness of the two endpoints (P_0, P_2) of QBC. Because these eight action parameters are learned by neural networks, the shapes and sizes of each stroke are different. The formula of QBC is defined as follows,

$$QBC(\alpha) = (1-\alpha)^2 P_0 + 2(1-\alpha)\alpha P_1 + \alpha^2 P_2, 0 \le \alpha \le 1. \tag{6}$$

The tangents to the QBC at P_0 and P_2 intersect at P_1. As α increases from 0 to 1, the curve starts from P_0 in the direction of P_1 and bends to end at P_2 from the direction of P_1. To further improve the accuracy, an action bundle strategy is adopted. The idea of action bundle is inspired by frame skip (Mnih et al. 2013), which is an important hyper-parameter in many RL tasks. Frame skip decides the granularity at which agents can observe the environment and select an action to perform. A parameter of frame skip K allows the agent to repeat a selected action at K frames. This strategy can explore the connection of similar states and save computing resources. Following the idea of frame skip, the connection is explored between different actions, which is called action bundle. In order to make actor can better explore the action space, actor picks out K actions from the action space to form an action bundle. Then, the segmentation executor performs K actions in one action bundle, which can further improve the accuracy of the segmentation result.

3.4 Modified experience replay memory for DDPG

The training samples in DRL algorithm are called transition. Each transition has five parameters, which are current state S, the selected action A based on S, instant reward R, next state S', and Terminal indicating whether the current state is terminated. Experience replay memory is used to store transition $(S, A, R, S', Terminal)$ and break the correlation between transitions by random sampling. When the ERM stores adequate number of samples by the interaction of agent with the environment, a mini-batch of transitions are random sampled from the memory for training agent. At each step, the action A and the state S are fed into critic for obtaining the long-term return Q. The accuracy of the critic affects both the ability of actor finding the best policy π and the efficiency of the algorithm. To improve the evaluation ability of the critic for the segmentation task, ground truth is added as a new parameter to the transition. Therefore, the new transition consists of $(S, A, R, S', GT, Terminal)$. Based on the new transition, the GT and S' are sent to critic for evaluation. The appearance of the ROI is usually similar to that of the surrounding tissue, which results an ambiguous boundary. In this situation, it is difficult for segmentation agent to understand the whole environment. Therefore, the ground truth is added in transition to help agent further understand the environment, which makes the segmentation more accurate. The modified ERM is presented in Fig. 3.

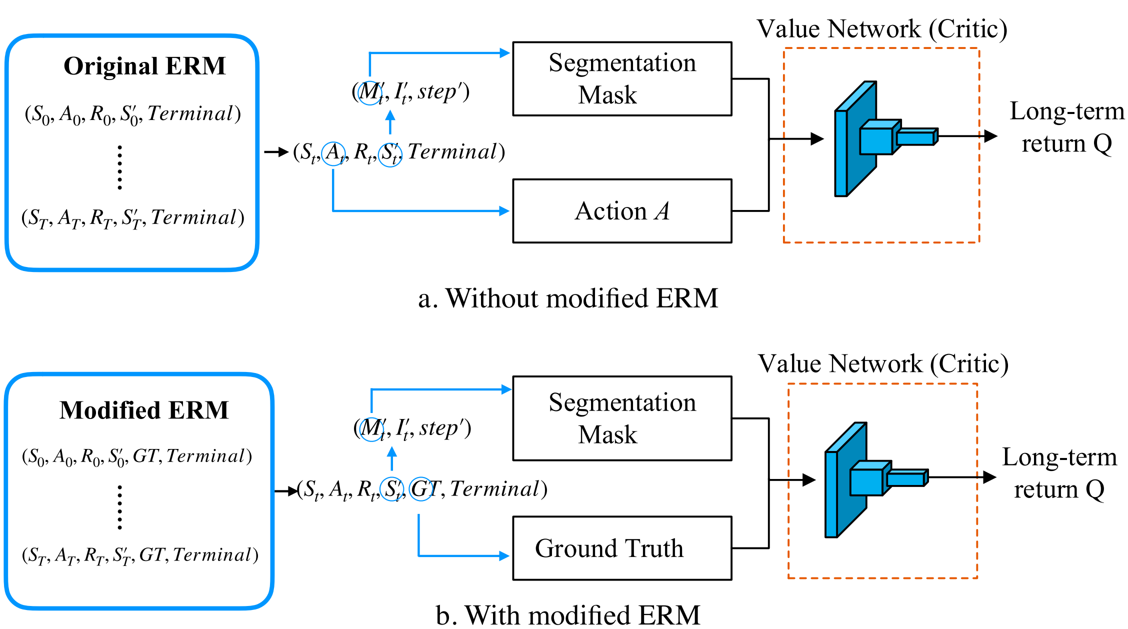


Fig. 3 The modified experience replay memory is used to change the input of critic for an accurate evaluation

4 Experiments

4.1 Data sets and evaluation metrics

Two types of medical image data sets were used for evaluation, which are prostate MR image data set and retinal fundus image data set. For the prostate MR image segmentation, the experiments were performed on three prostate MR image data sets, which contains 172 MR subjects. 142 subjects were used for training, which are from PROMISE12, ISBI2013, and in-house data sets. 30 subjects from PROMISE12 test data set are used for testing. All these images are fully labeled by the radiologists. For the retinal fundus image segmentation, REFUGE challenge dataset (Orlando et al. 2020) was used to evaluate the performance of the proposed method on multi-class data set. Two classes of ROI should be segmented from fundus images, which are optic cup and optic disc. This dataset consists of 400 training images, 400 validation images, and 400 testing images.

Four quantitative metrics were used for segmentation evaluation, which are Dice similarity coefficient (DSC), Hausdorff distance (HD), relative volume difference (RVD), and average boundary distance (ABD) (Tian et al. 2016). The DSC is defined as follows,

$$DSC = \frac{2|S_{gt} \cap S_m|}{|S_{gt}| + |S_m|} \times 100\%, \tag{7}$$

where $|S_{gt}|$ is the number of pixels of the ROI from the manually segmented ground truth. $|S_m|$ is the number of pixels of the ROI from the proposed method. DSC is a metric of area overlap between the predicted segmentation result and the ground truth. DSC values are expressed as a percentage varies from 0% (total mismatch) to 100% (perfect match).

A distance from a pixel x to a surface Y is defined as $d(x, Y) = min_{y\in Y}||x - y||$. The HD between two surfaces X and Y is calculated as,

$$HD(X, Y) = max[\max_{x\in X} d(x, Y), \max_{y\in Y} d(y, X)]. \tag{8}$$

Hausdorff distance measures the distance of the predicted segmentation and the ground truth. Smaller HD value means better performance of the segmentation methods.

The RVD evaluates the algorithm whether tends to over-segment or under-segment the ROI. The algorithm over-segments the ROI when RVD is negative, and vice versa. The relative volume difference is computed as follows:

$$RVD = 100 \times \left(\frac{|S_{gt}|}{|S_m|} - 1 \right). \tag{9}$$

The ABD is computed as follows:

$$\mathrm{ABD}\left(S_{gt}, S_m\right) = \frac{1}{N_{S_{gt}} + N_{S_m}} \left(\sum_{x\in S_{gt}} \min_{y\in S_m} \mathbf{d}(x, y) + \sum_{y\in S_m} \min_{x\in S_{gt}} \mathbf{d}(y, x) \right), \tag{10}$$

where $N_{S_{gt}}$ and N_{S_m} represent the number of pixels in the surface S_{gt} and S_m, respectively. $\mathbf{d}(x, y)$ denotes a distance from pixel x to pixel y.

4.2 Experiment details

In our experiments, the medical images were resized to 128×128 during training and testing. The segmentation agent was trained with Adam (Kingma and Ba 2014) for optimization. The mini-batch size was set as 64. All experiments were performed on a single NVIDIA GeForce RTX 2080Ti with 11G memory. The range of actor learning rate is [3e−4, 1e−4] and critic learning rate is [1e−3, 3e−4], which both decay every 800 training episodes. The reward discount factor γ is set as 0.955. The size of experience replay memory is set as 400. We set the action bundle $K = 5$ and step number $t = 3$.

4.3 Qualitative evaluation results

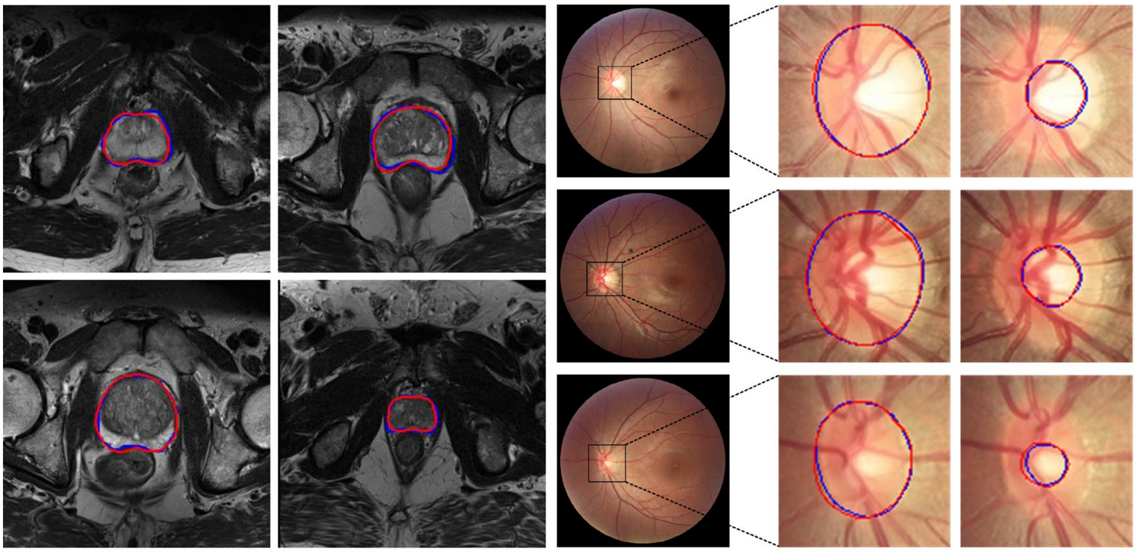
The performance of the proposed method was evaluated qualitatively by visualizing contours of the proposed method and the manually segmented ground truth. Figure 4 shows the qualitative results on prostate MR images and retinal fundus images. As shown in figure, the predicted contours (red curves) are very close to the ground truth (blue curves). Furthermore, the proposed method could achieve robust and accurate results across different subjects.

4.4 Quantitative results

4.4.1 Prostate MR dataset

Six state-of-the-art segmentation methods were adopted for evaluating the proposed method on prostate MR data set, which are Grab-Cut (Rother et al. 2004), PSPNet (Zhao et al. 2017), FCN (Long et al. 2015), U-Net (Ronneberger et al. 2015), V-Net (Milletari et al. 2016), and DeepLabV3+ (Chen et al. 2018). The comparison results are shown in Table 1. Four evaluation metrics were used in the experiment, which are region-based DSC and relative volume difference metrics, distance-based HD and average boundary distance metrics. The standard deviation of DSC and HD are also presented.

The proposed method could get a DSC of 93.69% ± 1.04%, a HD of 14.00 mm ± 6.37 mm, a RVD of − 0.30%, and an ABD of 1.8 mm for prostate MR data set. The results show that the proposed method achieves the highest DSC



(a) Prostate MR images **(b)** Retinal fundus images (optic disk and cup)

Fig. 4 The qualitative results of the proposed method on prostate MR images and retinal fundus images. The red curves represent the contours obtained by the proposed method, while the blue curves represent the ground truth

Table 1 Quantitative comparison between the proposed method and six segmentation methods

	DSC	Std. (DSC)	HD	Std. (HD)	RVD	ABD
Grab-Cut (Rother et al. 2004)	78.41	15.62	21.52	11.27	4.12	2.56
PSPNet (Zhao et al. 2017)	75.49	9.41	24.58	15.26	4.71	2.89
FCN (Long et al. 2015)	82.37	5.56	19.64	19.79	6.06	2.39
U-Net (Ronneberger et al. 2015)	84.71	6.52	15.92	6.85	2.40	1.89
V-Net (Milletari et al. 2016)	85.29	6.82	16.78	6.60	3.49	2.02
DeepLabV3+ (Chen et al. 2018)	86.45	5.09	23.08	19.07	– 6.18	2.20
Ours	**93.69**	**1.04**	**14.00**	**6.37**	**– 0.30**	**1.80**

Bold values indicate the best result

DSC (%), HD (mm), RVD (%), and ABD (mm)

value and the lowest HD in the quantitative comparison. In addition, the proposed method has the lowest standard deviation of both DSC and HD, which means that our method is robust to the different prostate MR volumes.

4.4.2 Retinal fundus dataset

REFUGE challenge dataset (Orlando et al. 2020) is used to further explore the effectiveness of the proposed method on multi-class segmentation. Two ROIs will be segmented, which are the optic cup and optic disc regions in the images. The optic disc, optic cup, and mean segmentation result are shown in Table 2.

From the Table 2, we can see that the proposed method gets the best performance for mean of cup and disk.

4.5 Alternative segmentation executors

The segmentation executor is trained based on a specific brushstroke. In the segmentation process, we tried three segmentation executors with different brushstroke shapes, which are triangle, round, and quadratic Bezier curve. The visualization of the segmentation executors is shown in Fig. 5.

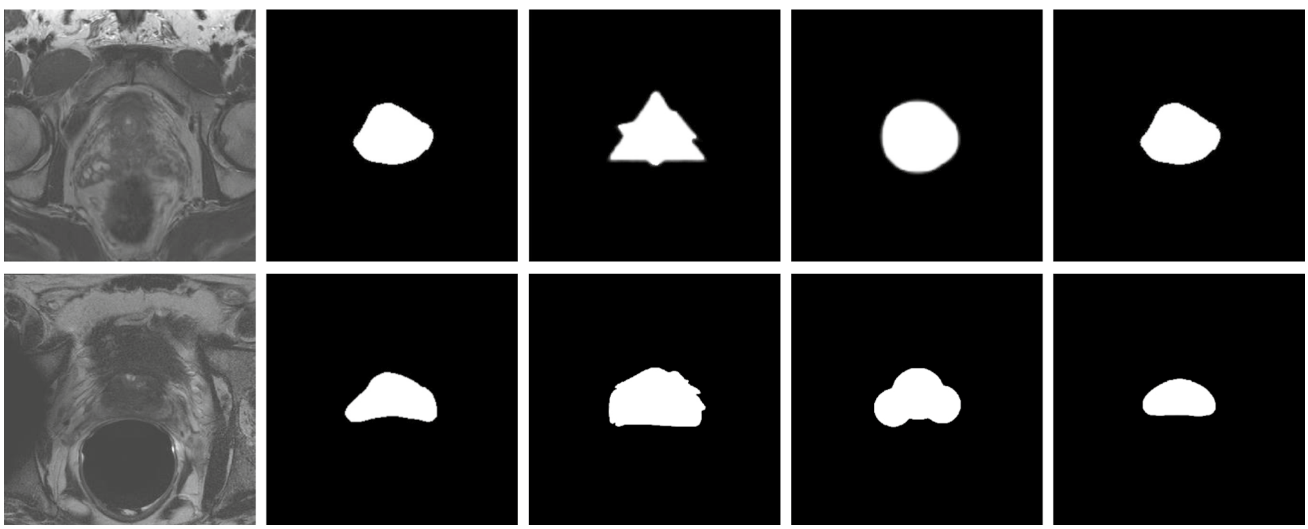
From the results of the segmentation, we can see that the QBC can get best results. We further compared the results obtained by applying these three different segmentation executors. The DSC results are shown in Fig. 6a.

As shown in Fig. 6a, QBC gets the highest DSC compared with triangle and round segmentation executors. In addition, all three segmentation executors get a satisfactory results at the third step. When the step further increases, the

Table 2 The optic disc, optic cup, and mean segmentation results on the fundus images

	PSPNet (Zhao et al. 2017)	FCN (Long et al. 2015)	U-Net (Ronneberger et al. 2015)	DeepLabv3+ (Chen et al. 2018)	Ours
Disc DSC (%)	88.13 ± 0.61	97.11 ± 0.14	97.72 ± 0.13	81.96 ± 0.26	97.29 ± 0.20
Disc HD (mm)	27.81 ± 1.26	9.30 ± 1.14	7.53 ± 1.87	36.14 ± 2.37	8.07 ± 1.40
Disc ABD (mm)	3.68 ± 0.31	1.27 ± 0.04	1.08 ± 0.04	14.50 ± 1.49	1.04 ± 0.05
Cup DSC (%)	67.65 ± 3.27	84.55 ± 1.36	85.68 ± 0.92	70.26 ± 1.82	93.15 ± 0.36
Cup HD (mm)	46.19 ± 3.86	15.10 ± 1.67	12.55 ± 2.10	21.53 ± 4.26	11.37 ± 1.75
Cup ABD (mm)	6.08 ± 0.71	2.46 ± 0.18	2.23 ± 0.14	3.62 ± 0.28	1.73 ± 0.11
Mean DSC (%)	77.89	90.83	91.70	76.11	**95.22**
Mean HD (mm)	37.00	12.20	10.04	28.84	**9.72**
Mean ABD (mm)	4.88	1.87	1.66	9.06	**1.38**

Bold values indicate the best result



(a) Prostate image **(b)** Ground truth **(c)** Triangle **(d)** Round **(e)** QBC

Fig. 5 The visualization of three segmentation executors on two prostate MR images.

DSC value has a very small change. Therefore, the proposed method only needs three steps to get the final segmentation results.

4.6 Ablation experiments

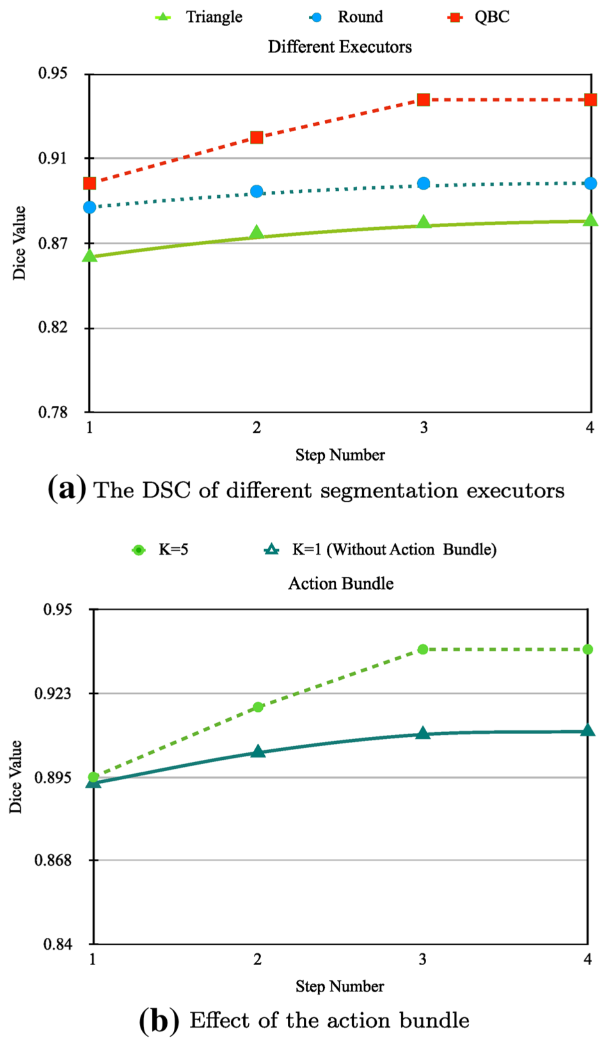
To further evaluate the effects of action bundle and the modified experience replay memory, two ablation experiments were performed. The first ablation experiment was performed to evaluate the effect of the action bundle. The results are shown in Fig. 6b. From the figure, we can see that the action bundle strategy could improve the accuracy.

We also adopt modified experience replay memory in our method for the second ablation experiment. The results are shown in Table 3. As shown in table, the use of action bundle and modified experience replay memory can both improve the segmentation performance.

4.7 Action bundle setting and time consumption

To evaluate the influence of K on segmentation performance, we set the K value to 1 (without action bundle), 3, 5, 7 for evaluation, respectively. The comparison results are shown in Table 4. From the table, we can see that the proposed method can get best result when K is set as 5.

At the same time, the time consumption was also evaluated. There are two factors affecting the efficiency of the proposed method, which are parameter K and the number of the step. When $K = 5$ and the step set as 3, the training time for each experiment is about an hour. It only spends



(a) The DSC of different segmentation executors

(b) Effect of the action bundle

Fig. 6 **a** It shows the DSC values of QBC segmentation executor indicated by red line. The DSC values of round segmentation executor are indicated by blue line, and the DSC values of triangle segmentation executor are indicated by green line. **b** The dash line represents the DSC value in different steps with action bundle. The solid line represents the DSC values without action bundle

Table 3 The DSC values of two ablation experiments

	Basic model	With modified ERM	With action bundle	With Both
DSC (%)	89.39	90.72	92.21	**93.69**
HD (mm)	15.28	14.87	14.21	**14.00**

Bold values indicate the best result

Table 4 The results of different action bundle settings

	$K=1$	$K=3$	$K=5$	$K=7$
DSC (%)	89.39	91.59	93.69	92.34
HD (mm)	15.28	14.69	14.00	14.12

0.08s for segmenting one medical image. Note that, There is no additional time consumption for action bundle strategy.

5 Conclusions

In this paper, we propose an automatic multi-step medical image segmentation method based on deep reinforcement learning algorithm. An agent is trained by deep deterministic policy gradient, which could segment ROIs from medical image in a multi-step manner. We adopt two strategies to further improve the accuracy of segmentation, which are action bundle and modified experience replay memory. Experimental results show that the proposed method could get state-of-the-art results. In the future, we will try to use this method for multi-organ semantic segmentation and other modalities images. The proposed method also can handle small particles in the segmentation task. Meanwhile, a certain amount of medical image segmentation tasks may have uncertain amount of separated regions. Depending on the generalization of the proposed method, it also can handle this situation. The proposed model can learn the ability of segmenting uncertain amount of separated regions. However, the segmentation accuracy depends on the complexity of the foreground regions.

Acknowledgements This work was supported in part by the National Natural Science Foundation of China under Grant no. 61876148. This work was also supported in part by the Fundamental Research Funds for the Central Universities no. XJJ2018254, and China Postdoctoral Science Foundation no. 2018M631164.

References

Aguirreramos H, Avinacervantes JG, Cruzaceves I, Ruizpinales J, Ledesma S (2018) Blood vessel segmentation in retinal fundus images using gabor filters, fractional derivatives, and expectation maximization. Appl Math Comput 339:568–587

Ahmadvand A, Daliri MR (2015) Improving the runtime of mrf based method for mri brain segmentation. Appl Math Comput 256:808–818

Alansary A, Oktay O, Li Y, Le Folgoc L, Hou B, Vaillant G, Kamnitsas K, Vlontzos A, Glocker B, Kainz B et al (2019) Evaluating reinforcement learning agents for anatomical landmark detection. Med Image Anal 53:156–164

Arulkumaran K, Deisenroth MP, Brundage M, Bharath AA (2017) Deep reinforcement learning: a brief survey. IEEE Signal Process Mag 34(6):26–38

Ba J, Mnih V, Kavukcuoglu K (2014) Multiple object recognition with visual attention. arXiv preprint arXiv:14127755

Caicedo JC, Lazebnik S (2015) Active object localization with deep reinforcement learning. In: Proceedings of the IEEE international conference on computer vision. IEEE, Santiago, pp 2488–2496

Castrejon L, Kundu K, Urtasun R, Fidler S (2017) Annotating object instances with a polygon-rnn. In: Proceedings of the IEEE conference on computer vision and pattern recognition. IEEE, Honolulu, pp 5230–5238

Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL (2017) Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans Pattern Anal Mach Intell 40(4):834–848

Chen LC, Zhu Y, Papandreou G, Schroff F, Adam H (2018) Encoder–decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV). Springer, Munich, pp 801–818

Dai J, He K, Sun J (2016) Instance-aware semantic segmentation via multi-task network cascades. In: Proceedings of the IEEE conference on computer vision and pattern recognition. IEEE, Las Vegas, pp 3150–3158

Eltanboly A, Ghazal M, Hajjdiab H, Shalaby A, Switala A, Mahmoud AM, Sahoo PK, Elazab MS, Elbaz A (2019) Level sets-based image segmentation approach using statistical shape priors. Appl Math Comput 340:164–179

Guotai W, Li Wenqi, Vercauteren T (2018) Interactive medical image segmentation using deep learning with image-specific fine tuning. IEEE Trans Med Imaging 37(7):1562–1573

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. IEEE, Las Vegas, pp 770–778

Huang Z, Heng W, Zhou S (2019) Stroke-based artistic rendering agent with deep reinforcement learning. arXiv preprint arXiv:190304411

Ioffe S, Szegedy C (2015) Batch normalization: Accelerating deep network training by internal covariate shift. arXiv preprint arXiv:150203167

Jagadeesan S, Subbiah J (2020) Real-time personalization and recommendation in adaptive learning management system. J Ambient Intell Humaniz Comput 2:1–11

Kim D, Lee T, Kim S, Lee B, Youn HY (2019) Adaptive packet scheduling in iot environment based on q-learning. J Ambient Intell Humaniz Comput 6:1–11

Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. arXiv preprint arXiv:14126980

Lillicrap TP, Hunt JJ, Pritzel A, Heess N, Erez T, Tassa Y, Silver D, Wierstra D (2015) Continuous control with deep reinforcement learning. arXiv preprint arXiv:150902971

Litjens GJS, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, Der Laak JAWMV, Van Ginneken B, Sanchez CI (2017) A survey on deep learning in medical image analysis. Med Image Anal 42:60–88

Liu F, Li S, Zhang L, Zhou C, Ye R, Wang Y, Lu J (2017) 3dcnn-dqn-rnn: a deep reinforcement learning framework for semantic parsing of large-scale 3d point clouds. In: Proceedings of the IEEE international conference on computer vision. IEEE, Venice, pp 5678–5687

Liu R, Lehman J, Molino P, Such FP, Frank E, Sergeev A, Yosinski J (2018) An intriguing failing of convolutional neural networks and the coordconv solution. In: Advances in neural information processing systems. MIT Press, Montreal, pp 9605–9616

Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. IEEE, Boston, pp 3431–3440

Lu H, Li B, Zhu J, Li Y, Li Y, Xu X, He L, Li X, Li J, Serikawa S (2017) Wound intensity correction and segmentation with convolutional neural networks, concurrency and computation: practice and experience. Concurr Comput Pract Exp 29(6):e3927

Lu H, Kondo M, Li Y, Tan J, Kim H (2020) Supervoxel graph cuts: an effective method for ggo candidate regions extraction on ct images. IEEE Consum Electron Mag 9(1):61–66

Madani Y, Ezzikouri H, Erritali M, Hssina B (2019) Finding optimal pedagogical content in an adaptive e-learning platform using a new recommendation approach and reinforcement learning. J Ambient Intell Humaniz Comput 12:1–16

Milletari F, Navab N, Ahmadi SA (2016) V-net: fully convolutional neural networks for volumetric medical image segmentation. In: International conference on 3D vision (3DV). IEEE, Stanford, pp 565–571

Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, Graves A, Riedmiller M, Fidjeland AK, Ostrovski G et al (2015) Human-level control through deep reinforcement learning. Nature 518(7540):529

Mnih V, Badia AP, Mirza M, Graves A, Lillicrap T, Harley T, Silver D, Kavukcuoglu K (2016) Asynchronous methods for deep reinforcement learning. In: International conference on machine learning. ACM, New York, pp 1928–1937

Mnih V, Kavukcuoglu K, Silver D, Graves A, Antonoglou I, Wierstra D, Riedmiller M (2013) Playing atari with deep reinforcement learning. arXiv preprint arXiv:13125602

Orlando JI, Fu H, Breda JB, van Keer K, Bathula DR, Diaz-Pinto A, Fang R, Heng PA, Kim J, Lee J et al (2020) Refuge challenge: a unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. Med Image Anal 59:101570

Rao Y, Lu J, Zhou J (2017) Attention-aware deep reinforcement learning for video face recognition. In: Proceedings of the IEEE international conference on computer vision. IEEE, Venice, pp 3931–3940

Ronneberger O, Fischer P, Brox T (2015) U-net: Convolutional networks for biomedical image segmentation. In: International conference on medical image computing and computer-assisted intervention. Springer, Munich, pp 234–241

Rother C, Kolmogorov V, Blake A (2004) Grabcut: Interactive foreground extraction using iterated graph cuts. ACM Trans Graph 23(3):309–314

Sahba F, Tizhoosh HR, Salama MM (2008) Application of reinforcement learning for segmentation of transrectal ultrasound images. BMC Med Imaging 8(1):8

Salimans T, Kingma DP (2016) Weight normalization: a simple reparameterization to accelerate training of deep neural networks. In: Advances in neural information processing systems. MIT Press, Barcelona, pp 901–909

Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O (2017) Proximal policy optimization algorithms. arXiv preprint arXiv:170706347

Shi W, Caballero J, Huszar F, Totz J, Aitken AP, Bishop R, Rueckert D, Wang Z (2016) Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. IEEE, Las Vegas, pp 1874–1883

Silver D, Lever G, Heess N, Degris T, Wierstra D, Riedmiller M (2014) Deterministic policy gradient algorithms. In: International conference on machine learning. ACM, Beijing, pp 1–9

Silver D, Huang A, Maddison CJ, Guez A, Sifre L, Van Den Driessche G, Schrittwieser J, Antonoglou I, Panneershelvam V, Lanctot M et al (2016) Mastering the game of go with deep neural networks and tree search. Nature 529(7587):484–489

Song G, Myeong H, Mu Lee K (2018) Seednet: automatic seed generation with deep reinforcement learning for robust interactive segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. IEEE, Salt Lake City, pp 1760–1768

Sutton RS, McAllester DA, Singh SP, Mansour Y (2000) Policy gradient methods for reinforcement learning with function approximation. In: Advances in neural information processing systems. ACM, New York, pp 1057–1063

Tian Z, Liu L, Zhang Z, Fei B (2016) Superpixel-based segmentation for 3d prostate mr images. IEEE Trans Med Imaging 35(3):791–801

Wang Z, Sarcar S, Liu J, Zheng Y, Ren X (2018) Outline objects using deep reinforcement learning. arXiv preprint arXiv:180404603

Watkins CJ, Dayan P (1992) Q-learning. Mach Learn 8(3–4):279–292

Wu J, Li G, Lu H, Kim H (2019) Multi-organ segmentation from abdominal ct with random forest based statistical shape model. In: International conference on biomedical signal and image processing. ACM, New York, pp 67–70

Xiang S, Li H (2017) On the effects of batch and weight normalization in generative adversarial networks. arXiv preprint arXiv:17040 3971

Xie N, Zhao T, Tian F, Zhang XH, Sugiyama M (2015) Stroke-based stylization learning and rendering with inverse reinforcement learning. In: Twenty-fourth international joint conference on artificial intelligence. AAAI, Palo Alto, pp 2531–2537

Xu X, Lu H, Song J, Yang Y, Shen HT, Li X (2019) Ternary adversarial networks with self-supervision for zero-shot cross-modal retrieval. In: IEEE Trans Cybern. https://doi.org/10.1109/TCYB.2019.2928180

Yoshino Y, Miyajima T, Lu H (2017) Automatic classification of lung nodules on mdct images with the temporal subtraction technique. Int J Comput Assist Radiol Surg 12:1789–1798

Zhao H, Shi J, Qi X, Wang X, Jia J (2017) Pyramid scene parsing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. IEEE, Honolulu, pp 2881–2890

Zhu Y, Zhao D (2019) Vision-based control in the open racing car simulator with deep and reinforcement learning. J Ambient Int Humaniz Comput 9:1–13

Zhu Y, Mottaghi R, Kolve E, Lim JJ, Gupta A, Fei-Fei L, Farhadi A (2017) Target-driven visual navigation in indoor scenes using deep reinforcement learning. In: 2017 IEEE international conference on robotics and automation (ICRA). IEEE, Singapore, pp 3357–3364