Medical Image Analysis (2022)

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Deep reinforcement learning in medical imaging: A literature review

S. Kevin Zhou[a,*], Hoang Ngan Le[b,*], Khoa Luu[b], Hien V. Nguyen[c], Nicholas Ayache[d]

[a]Medical Imaging, Robotics, and Analytic Computing Laboratory and Engineering (MIRACLE)
School of Biomedical Engineering & Suzhou Institute for Advanced Research, University of Science and Technology of China, Suzhou 215123, China;
Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190 China
[b]CSCE Department, University of Arkansas, Fayetteville, AR 72701
[c]ECE Department, University of Houston, Houston, TX 77204
[d]INRIA, Sophia Antipolis-Mediterranean Centre, France

ARTICLE INFO

Article history:

Keywords: Deep reinforcement learning, Medical imaging, Survey

ABSTRACT

Deep reinforcement learning (DRL) augments the reinforcement learning framework, which learns a sequence of actions that maximizes the expected reward, with the representative power of deep neural networks. Recent works have demonstrated the great potential of DRL in medicine and healthcare.This paper presents a literature review of DRL in medical imaging. We start with a comprehensive tutorial of DRL, including the latest model-free and model-based algorithms. We then cover existing DRL applications for medical imaging, which are roughly divided into three main categories: (i) parametric medical image analysis tasks including landmark detection, object/lesion detection, registration, and view plane localization; (ii) solving optimization tasks including hyperparameter tuning, selecting augmentation strategies, and neural architecture search; and (iii) miscellaneous applications including surgical gesture segmentation, personalized mobile health intervention, and computational model personalization. The paper concludes with discussions of future perspectives.

1. Introduction

Reinforcement learning is a framework for learning a sequence of actions that maximizes the expected reward Sutton and Barto (2018); Li (2017). Deep reinforcement learning (DRL) is the result of marrying deep learning with reinforcement learning Mnih et al. (2013). DRL allows reinforcement learning to scale up to previously intractable problems. Deep learning and reinforcement learning were selected by MIT Technol-

*Corresponding authors.

ogy Review as one of 10 Breakthrough Technologies[1] in 2013 and 2017, respectively. The combination of these two powerful technologies currently constitutes one of the state-of-the-art frameworks in artificial intelligence.

Recent years have witnessed rapid progress in DRL, resulting in significant performance improvement in many areas, including games Mnih et al. (2013), robotics Finn et al. (2016a), natural language processing Luketina et al. (2019), and computer vision Bernstein and Burnaev (2018). Unlike supervised learning, DRL framework can deal with sequential decisions, and learn with highly delayed supervised information (e.g., success or failure of the decision is available only after multiple time steps). Since the DRL agent's decisions affect the world state, one cannot propagate the gradient from the reward to past actions without explicitly modeling the joint distribution between decisions and the world state. Traditional supervised learning lacks this explicit modeling, thus, is not effective in learning with sequential actions and delayed rewards. DRL can also deal with *non-differentiable metrics*. For example, one can use DRL to search for an optimal deep network architecture Zoph and Le (2016) or parameter settings to maximize the classification accuracy, which is clearly non-differentiable with respect to the number of layers or the choice of non-linear rectifier functions. However, DRL is not the only viable approach for network architecture search (NAS). In Liu et al. (2018), network architecture is sought using supervised learning. Another use of DRL is in finding efficient search sequence for speeding up detection Ghesu et al. (2017) or optimal transformation sequence for improving registration accuracy. DRL can also mitigate the issue of high memory consumption in processing high-dimensional medical images. For example, a DRL-based object detection can focus on a small image region at a time, which incurs a lower memory footprint, then decide next regions to process.

Despite its successes, application of this DRL technology to medical imaging remains to be fully explored Zhou et al. (2020). This is partly due to the lack of a systematic understanding of the DRL's strengths or weaknesses when applying to medical data. To this end, we organized a MICCAI 2018 tutorial [2], with its goal of bridging the gap by providing a comprehensive introduction to deep reinforcement learning methods in terms of theories, practice, and future directions. The tutorial contained multiple presentations from active researchers in DRL, covering state-of-the-art and explaining in-depth how DRL was applied in a selected set of topics such as neural architecture search Zoph and Le (2016), detection Ghesu et al. (2016), segmentation Sahba et al. (2006), and controlling of surgical robots Liu and Jiang (2018). This tutorial forms the basis of the paper. However, in this paper we go much beyond the tutorial and expand it with many state-of-the-art contents.

Our goal is to provide our readers good knowledge about of the principle of DRL and a thorough coverage of the latest examples of how DRL is used for solving medical imaging tasks. We structure the rest of paper as follows: (i) introduction to deep reinforcement learning with its generation framework and latest learning strategies; (ii) how to use DRL for solving medical image analysis tasks, which is the main

[1] https://www.technologyreview.com/10-breakthrough-technologies/

[2] The tutorial is available online at https://www.hvnguyen.com/deepreinforcementlearning

part that covers the literature review; (iii) fundamental challenges and future potential of DRL in medical domains; and (iv) conclusions.

There are a few DRL survey papers such as Arulkumaran et al. (2017); Li (2017); François-Lavet et al. (2018). However, they cover basic principles and various applications of DRL. Our survey centers around the essential topic of DRL in medical imaging and marginally in healthcare applications.

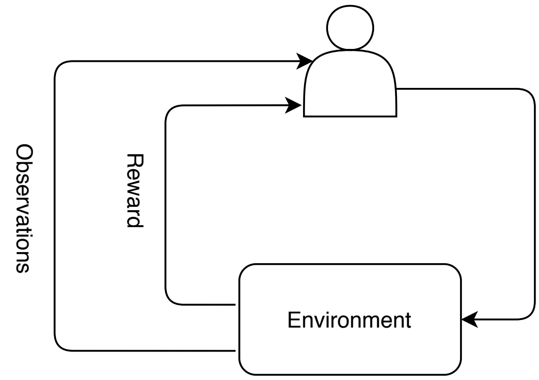

Fig. 1. An illustration of agent-environment interaction in RL.

2. Basics of Reinforcement Learning

Here we focus on how the RL problem can be formalized as an agent that is able to make decisions in an environment to optimize some objectives. Key aspects of RL include: (i) Addressing the sequential decision making; (ii) There is no supervisor, only a reward presented as a scalar number; (iii) Feedback is highly delayed. The interaction between agent and environment is illustrated in Fig. 1. The standard theory of RL is defined by a Markov Decision Process (MDP) in which rewards depend on the last state and action only. However, most of real-world decision-making is based non-Markovian model in which the next state depends on more than the current state and action. This work only focuses on MDP, however, the readers can learn more of the recent research on non-MDP in Clarke et al. (2015), Gaon and Brafman (2020), Majeed and Hutter (2018) with different aspects discussed i.e. non-MDP using Spiking Neural Networks (SNNs), Non-Markovian Rewards and Q-learning convergence for Non-MDP.

2.1. Markov Decision Process

An MDP is typically defined by five elements $\{S, A, T, R, \gamma\}$; where S is a set of *state*/observation space of an environment and s_0 is a starting state; A is set of *actions* the agent can choose from; T is a *transition probability* function $T(s_{t+1}|s_t, a_t)$, specifying the probability that the environment will transition to state $s_{t+1} \in S$ if the agent takes action $a \in \mathcal{A}$ in state $s \in S$; R is a *reward* function where $r_{t+1} = R(s_t, s_{t+1})$ is a reward received for taking action a_t at state s_t and transfer to the next state s_{t+1}; $\gamma \in [0, 1]$ is a discount factor which determines how much a agent cares about rewards in the future. A full sequence $(s_0, a_0, r_1, s_1, a_1, r_2, ...)$ is called a *trajectory* $\mathcal{T}$. Theoretically, a trajectory goes to infinity, but the episodic property holds in most practical cases. One trajectory of some finite length τ, is called an *episode*.

In order to estimate how good it is for an agent to utilize policy π to visit state s, a value function is introduced. The value is the mathematical expectation of return and value approximation is obtained by Bellman expectation equation as follows:

$$V^{\pi}(s_t) = \mathbb{E}[r_{t+1} + \gamma V^{\pi}(s_{t+1})]. \quad (1)$$

$V^{\pi}(s_t)$ is also known as the *state-value function* as it evaluates the value of a state s at time step t, and the expectation term can be expanded as a product of policy, transition probability, and return as follows:

$$V^{\pi}(s_t) = \sum_{a_t \in A} \pi(a_t|s_t) \sum_{s_{t+1} \in S} T(s_{t+1}|s_t, a_t)[R(s_t, s_{t+1}) + \gamma V^{\pi}(s_{t+1})]. \tag{2}$$

This equation Eq.2 is called a Bellman equation Li (2017); Geist and Pietquin (2010); Lagoudakis (2017). The goal of an MDP problem is to compute an optimal policy π^* such that $V^{\pi^*}(s) > V^{\pi}(s)$ for every policy π and every state $s \in S$. To represent the optimal value of each state-action, Q-value is defined as

$$Q^{\pi^*}(s_t, a_t) = \sum_{s_{t+1}} T(s_{t+1}|s_t, a_t)[R(s_t, s_{t+1}) + \gamma V^{\pi^*}(s_{t+1})]. \tag{3}$$

2.2. MDP Solutions

Many solution techniques are available to compute an optimal policy for a given MDP. In general, these techniques can be divided into model-free and model-based methods, depending on whether an explicit model is constructed or not. Here, "model" refers to the environment itself that is defined by the two quantities: transition probability function $T(s_{t+1}|s_t, a_t)$ and reward function $R(s_t, s_{t+1})$.

2.2.1. Model-based methods

Such methods exploit learned or given world dynamics, *i.e.*, $T(s_{t+1}|s_t, a_t)$, $R(s_t, s_{t+1})$. There are four main model-based techniques as follows:

Value Iteration: The objective of value function methods is to obtain the best policy by maximizing the value functions in each state. Value iteration specifies the optimal policy by iterating the Bellman updating.

Transition models: Transition models decide how to map from a state s, taking action a to the next state (s') and it strongly affects the performance of model-based RL algorithms. Depending on whether predicting the future state s' is based on the probability distribution of a random variable or not, there are two main approaches in this group: stochastic and deterministic.

Policy search: Policy search approach directly searches for the optimal policy by modifying its parameters whereas the value function methods indirectly find the actions that maximize the value function at each state.

Return functions: A return function decides how to aggregate rewards or punishments over an episode. It affects both the convergence and the feasibility of the model.

In practice, transition and reward functions are rarely known and hard to model. The comparative performances among all model-based techniques are reported in Wang et al. (2019) with over 18 benchmarking environments including noisy ones.

2.2.2. Model-free methods:

Such methods learn through the experiences gained from interactions with the environment, that is, a model-free method tries to estimate the transition probability function and the reward function from the experiences to exploit them in acquisition of policy. Policy-based and value-based algorithms are popularly used in model-free methods.

Policy-based: In this approach, RL task is considered as optimization with stochastic first-order optimization. Policy gradient methods directly optimize the discounted expected reward, *i.e.*, $\mathcal{G}(\pi) \rightarrow \max_{\pi}$, to obtain the optimal policy π^* without any additional information about MDP. To do so, approximate estimations of gradient with respect to policy parameters are used. Taking Williams (1992) as an example, policy gradient parameterizes the policy and updates parameters θ:

$$\mathcal{G}^{\theta}(\pi) = \mathbb{E}_{\mathcal{T}_\tau} \sum_{t=0} log(\pi_\theta(a_t|s_t))\gamma^t \mathcal{R}, \quad (4)$$

where $\mathcal{R}$ is the total accumulated return.

Value-based: In this approach, the optimal policy π^* is implicitly conducted by gaining an approximation of optimal Q-function $Q^*(s,a)$. In value-based methods, agents update the value function to learn suitable policy while policy-based RL agents learn the policy directly. Q-learning is a typical value-based method. The updating rule of Q-learning with a learning rate λ is defined as:

$$Q(s_t, a_t) = Q(s_t, a_t) + \lambda\delta_t, \quad (5)$$

where $\delta_t = R(s_t, s_{t+1}) + \gamma \arg\max_a Q(s_{t+1}, a) - Q(s_t, a)$ is the temporal difference (TD) error.

Actor-critic: While value-based methods such as Q-learning suffer from poor convergence, policy-based methods tend to converge to local maximas and suffer from high variance, actor-critic methods address the aforementioned limitations. While an actor controls how an agent behaves, a critic measures how good the taken action is. Actor-critic is an improvement of policy gradient with a value-based critic Γ and is rewritten as:

$$\mathcal{G}^{\theta}(\pi) = \mathbb{E}_{\mathcal{T}_\tau} \sum_{t=0} log(\pi_\theta(a_t|s_t))\gamma^t \Gamma_t. \quad (6)$$

The critic function Γ can be defined as $Q^\pi(s_t, a_t)$ or $Q^\pi(s_t, a_t) - V^\pi_t$ or $R[s_{t-1}, s_t] + V^\pi_{t+1} - V^\pi_t$.

REINFORCE. REINFORCE was introduced by Williams (1992) to approximately calculate the gradient in Eq. (13) by using Monte-Carlo estimation. In REINFORCE approximate estimator, Eq. (13) is reformulated as:

$$\nabla_\theta \mathcal{G}(\theta) \approx \sum_{\mathcal{T}}^{N} \sum_{t=0} \gamma^t \nabla_\theta \log\pi_\theta(a_t|s_t)(\sum_{t'=t} \gamma^{t'-t} R(s_{t'}, s_{t'+1})), \quad (7)$$

where $\mathcal{T}$ is trajectory distribution. Theoretically, REINFORCE can be straightforwardly applied into any parametric $\pi_\theta(a|s)$. However, it is impractical to use it because it is time consuming for convergence and there are local optima. Based on the observation that the convergence rate of stochastic gradient descent directly depends on the variance of gradient estimation, variance-reducing technique was proposed to address naive REINFORCE's limitations by adding a term that reduces the variance without affecting the expectation.

Factors	Model-based RL	Model-free RL
#iterations between agent & environment	Small	Big
Convergence	Fast	Slow
Prior knowledge of transitions	Yes	No
Flexibility	Strongly depends on a learnt model	Adjust based on trials and errors

Table 1. Comparison between model-based RL and model-free RL

Figure 2.2.2 summarizes different RL approaches. The comparison between model-based RL and model-free RL is given in Table 1. In this Table, we compare different factors that invoke the effectiveness of a RL method where it is model-based or model-free.

3. Introduction to Deep Reinforcement Learning

Thanks to the rich context representation of Deep Learning (DL), DRL was proposed as a combination of RL and DL and has achieved rapid developments. Under DRL, the aforementioned value and policy can be expressed by a neural network, which allows to deal with a continuous state or action Lee et al. (2018); Masson et al. (2016) that is hard for a table represen-

- Reinforcement Learning
  - Model-based
    - Value Functions
      - Dynamic Programming (DP) Levine and Koltun (2014), Morimoto et al. (2003)
      - Temporal Difference (TD) Martinez-Marin and Duckett (2005)
      - Monte Carlo Hester et al. (2011)
    - Transition Models
      - Decision Tree Nguyen et al. (2013)
      - Linear Regression Mordatch et al. (2016)
      - Gaussian Process Deisenroth et al. (2014), Kupcsik et al. (2017), Andersson et al. (2015)
      - Expectation Maximization Coates et al. (2009)
      - Dynamic Bayesian Nguyen et al. (2013)
    - Policy Search
      - Gradient-based El-Fakdi and Carreras (2008), Morimoto and Atkeson (2009)
      - Information theory Kupcsik et al. (2017), Kupcsik et al. (2013)
      - Sampling based Bagnell and Schneider (2001)
    - Return Functions
      - Discounted returns functions Bagnell and Schneider (2001), Depraetere et al. (2014), Wilson et al. (2014)
      - Averaged returns functions Boedecker et al. (2014), Abbeel et al. (2010)
  - Model-free
    - Policy-based Kober and Peters (2014), REINFORCE Williams (1992)
    - Value-based
      - Online TD Rummery and Niranjan (1994)
      - Off-policy TD Watkins and Dayan (1992)
    - Actor-critic Peters and Schaal (2008)

Fig. 2. Summary of RL approaches with model-based and model-free techniques.

tation. Similar to RL, DRL can be categorized into model-based algorithms 3.2 and model-free algorithms 3.1 which will be introduced in this section.

3.1. Model-free DRL algorithms

There are three approaches, namely, value-based DRL methods, policy gradient DRL methods and actor-critic DRL methods to implement model-free algorithms. The three approaches are detailed as follows.

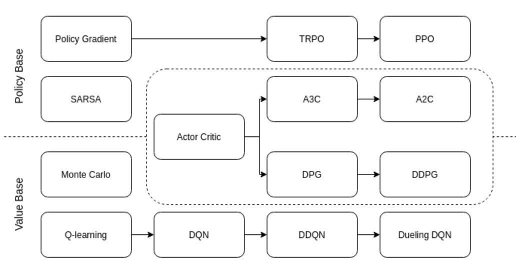

Fig. 3. A roadmap of model-free reinforcement learning algorithms.

3.1.1. Value-based DRL methods

Deep Q-Learning Network (DQN). DQN Mnih et al. (2015) is the most famous DRL model which learns policies directly from high-dimensional inputs by a deep neural network as given in Fig. 4(a). In general, QDN stabilizes the learning of Q-function by experience replay and the frozen target network. In experience replay, the agent's experiences at each time step are stored in a replay buffer (a fixed size dataset) and the Q-network is updated by SGD with sampled from minibatch data. Compared to standard online Q-learning, experience replay aims to avoid divergence, remove sample correlations, improve data efficiency. However, replay buffer does not differentiate important transitions. Target network addresses this limitation by proposing another network during Q-learning update process.

Taking the regression problem as an instance and letting y denote the target of our regression task, the regression with input (s, a), target $y(s, a)$ and the MSE loss function. The output y and MSE loss are defined as in Eq.(8).

$$\begin{aligned} y(s_t, a_t) &= R(s_t, s_{t+1}) + \gamma \max_{a_{t+1}} Q^*(s_{t_1}, a_{t+1}, \theta_t); \\ \mathcal{L}^{DQN} &= \mathcal{L}(y(s_t, a_t), Q^*(s_t, a_t, \theta_t)) \\ &= \|y(s_t, a_t) - Q^*(s_t, a_t, \theta_t)\|^2; \end{aligned} \tag{8}$$

where θ is vector of parameters, $\theta \in \mathbb{R}^{|S||R|}$ and s_{t+1} is a sample from $T(s_{t+1}|s_t, a_t)$ with input of (s_t, a_t). Q^* is Q-value under the optimal policy π^*.

Minimizing the loss function yields a gradient descent step formula to update θ as follows:

$$\theta_{t+1} = \theta_t - \alpha_t \frac{\partial \mathcal{L}^{DQN}}{\partial \theta}, \tag{9}$$

where α_t is a learning rate.

Double DQN. An improvement of DQN was introduced by Double DQN van Hasselt et al. (2015). One of the main limitation of DQN is that the values of Q^* are tend to overestimation because of *max* in Eq. (8), $y(s, a) = R(s, s') + \gamma \max_{a'} Q^*(s', a', \theta)$ shifts Q-value estimation towards either to the actions with high reward or to the actions with overestimating approximation error. Double DQN is an improvement of DQN by combining double Q-learning Hasselt (2010) with DQN to reduce observed overestimations with better performance.

The easiest but most expensive implementation of double DQN is to run two independent DQNs as follows:

$$\begin{aligned} y_1 &= R(s_t, s_{t+1}) + \gamma Q_1^*(s_{t+1}, \arg\max_{a_{t+1}} Q_2^*(s_{t+1}, a_{t+1}; \theta_2); \theta_1), \\ y_2 &= R(s_t, s_{t+1}) + \gamma Q_2^*(s_{t+1}, \arg\max_{a_{t+1}} Q_1^*(s_{t+1}, a_{t+1}; \theta_1); \theta_2). \end{aligned} \tag{10}$$

In Eq. 10, the current Q is used to select actions, and the target Q^* is used to evaluate actions.

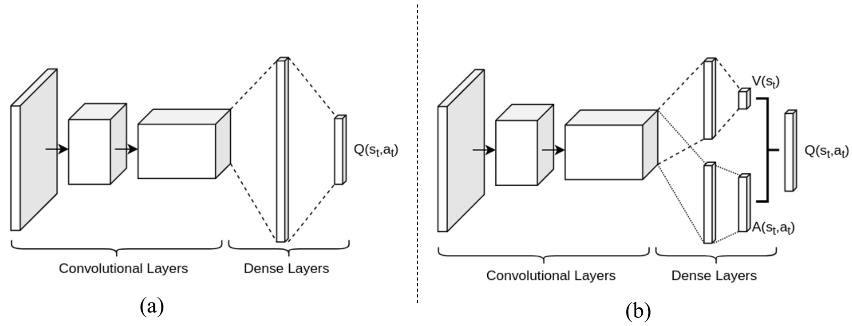

Fig. 4. (a): Network structure of Deep Q-Network (DQN), where Q-values Q(s,a) are generated for all actions for a given state. (b): Network structure of Dueling DQN, where value function $V(s)$ and advantage function $A(s,a)$ are combined to predict Q-values $Q(s,a)$ for all actions for a given state.

Dueling DQN. In DQN, when the agent visits unfavorable state, instead of lowering its value V^*, it remembers only low pay-off by updating Q^*. In order to address this limitation, Dueling DQN Wang et al. (2015) incorporates approximation of V^* explicitly in computational graph by introducing an advantage function as follows:

$$A^{\pi}(s_t, a_t) = Q^{\pi}(s_t, a_t) - V^{\pi}(s_t). \tag{11}$$

Therefore, Q-value is rewritten as

$$Q^*(s, a) = A^*(s, a) + V^*(s)$$

This implies that the feature map from DL is decomposed into with two parts corresponding to $V^*(v)$ and $A^*(s, a)$ as illustrated in Fig.4(b). In practical implementation, Dueling DQN is formulated as follows:

$$Q^*(s_t, a_t) = V^*(s_t) + A^*(s_t, a_t) - mean_{a_{t+1}} A^*(s_t, a_{t+1}).$$

Furthermore, to address the limitation of memory and imperfect information at each decision point, Deep Recurrent Q-Network (DRQN) Graves et al. (2013) employed RNN into DQN by replacing the first fully-connected layer with an RNN. Multi-step DQN De Asis et al. (2018) is one of the most popular improvements of DQN by substituting one-step approximation with N-steps.

3.1.2. *Policy gradient DRL methods*

Policy gradient theorem. Different from value-based DRL methods, policy gradient DRL optimizes the policy directly by optimizing the following objective function which is defined as a function of θ:

$$\mathcal{G}(\theta) = \mathbb{E}_{\mathcal{T} \sim \pi_\theta} \sum_{t=1} \gamma^{t-1} R(s_{t-1}, s_t) \rightarrow \max_{\theta}. \tag{12}$$

For any MDP and differentiable policy π_θ, the gradient of objective Eq. (12) is defined by policy gradient theorem Sutton et al. (2000) as follows:

$$\nabla_\theta \mathcal{G}(\theta) = \mathbb{E}_{\mathcal{T} \sim \pi_\theta} \sum_{t=0} \gamma^t Q^{\pi}(s_t, a_t) \nabla_\theta \log \pi_\theta(a_t|s_t). \tag{13}$$

3.1.3. *Actor-critic DRL algorithm*

Compared with value-based methods, policy gradient methods are better for continuous and stochastic environments and have a faster convergence Peters and Bagnell (2010); Lee et al. (2018). However, value-based methods are more sample efficient and steady Nachum et al. (2017); Liu et al. (2020c). Lately, actor-critics Konda and Tsitsiklis (2000) Mnih et al. (2016) was invented to take advantages from both value-based and policy gradient while limiting their drawbacks. Actor-critic architecture computes the policy gradient using a value-based critic function to estimate expected future reward. The principal idea of actor-critics is to divide the model into two parts: (i) computing an action based on a state and (ii) producing the Q value of the action. As given in Fig. 5, the actor takes as input the state s_t and outputs the best action a_t. It essentially controls how the agent behaves by learning the optimal policy (policy-based). The critic, on the other hand, evaluates the action by computing the value function (value based). The most basic actor-critic method (beyond the tabular case) is naive policy gradients (REINFORCE). The relationship between actor-critic is compared as a kid-mom relationship. The kid/actor explores the environment around with new actions while the mom/critic watches the kid and criticize/compliments. The kid then adjusts his behavior based on what his mom tells him. When the kid gets older, he is able to realize which action is bad/good.

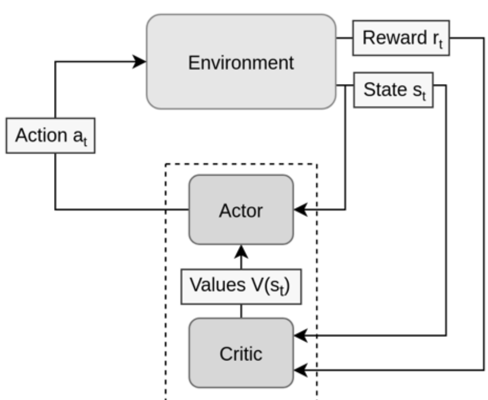

Fig. 5. Flowchart showing the structure of actor critic algorithm. Action a, state s, reward r

Advantage actor-critic (A2C). Advantage actor-critic (A2C) Mnih et al. (2016) consists of two neural networks, *i.e.*, an actor network $\pi_\theta(a_t|s_t)$ representing for policy and a critic network V^π_θ with parameters θ approximately estimating actor's performance.

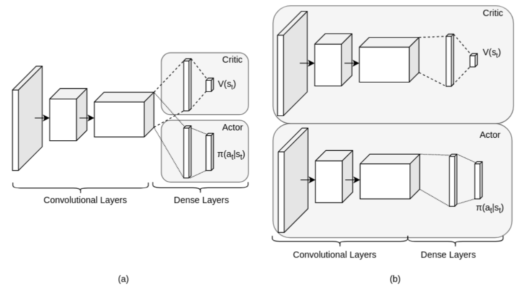

Fig. 6. An illustration of Actor-Critic algorithm in two cases: sharing parameters (a) and not sharing parameters (b).

At time step t, the A2C algorithm can be implemented as following steps:

- Step 1: Compute advantage function:

$$A^\pi(s_t, a_t) = R(s_t, s_{t+1}) + \gamma V^\pi_\theta(s_{t+1}) - V^\pi_\theta(s_t) \quad (14)$$

- Step 2: Compute target value:

$$y = R(s_t, s_{t+1}) + \gamma V^\pi_\theta(s_{t+1}) \quad (15)$$

- Step 3: Compute critic loss with MSE loss:

$$\mathcal{L} = \frac{1}{B}\sum_T ||y - V^\pi(s_t))||^2 \quad (16)$$

, where B is batch size and $V^\pi(s_t)$ is defined by:

$$V^\pi(s_t) = \mathbb{E}_{a_t\sim\pi(a_t|s_t)}\mathbb{E}_{s_{t+1}\sim T(s_{t+1}|a_t,s_t)}(R(s_t, s_{t+1}) + \gamma V^\pi(s_{t+1})) \quad (17)$$

- Step 4: Compute critic gradient:

$$\triangledown^{critic} = \frac{\partial\mathcal{L}}{\partial\theta} \quad (18)$$

- Step 5: Compute actor gradient:

$$\triangledown^{actor} = \frac{1}{B}\sum_T \triangledown_\theta \log\pi(a_t|s_t)A^\pi(s_t, a_t) \quad (19)$$

Asynchronous advantage actor critic (A3C). Besides A2C, asynchronous advantage actor critic (A3C) Mnih et al. (2016) is another strategy to implement an actor critic agent. To meet memory efficiency, A3C asynchronously executes different agents in parallel on multiple instances of the environment instead of experience replay as in A2C. Because of the asynchronous nature of A3C, some worker works with older values of the parameters and hence the aggregating update is not optimal. On the other hand, A2C synchronously updates the global network. A2C waits until all workers finished their training and calculated their gradients to average them, to update the global network.

In order to overcome the limitation of speed, Babaeizadeh et al. (2016) proposed GA3C which achieves a significant speed up compared to the original CPU implementation. To more effectively train A3C, Holliday and Le (2020) proposed FFE which forces random exploration at the right time during a training episode, which leads to improved training performance.

The structure of an actor-critic algorithm can be divided into two types, depending on whether or not parameter sharing is involved, as illustrated in Fig.6 in two cases of sharing network parameters and not sharing network parameters. In the first case, both actor and critic are trained on the same network architecture whereas the loss functions of each are defined independently via different FC layers. In the other case, the actor and critic are separately trained on two different networks. Compared to the first case, the second case may provides between performance but the network complexity is more expensive.

3.2. Model-based algorithms

We have discussed so far model-free methods including the value-based approach and policy gradient approach. In this section, we focus on the model-based approach, which deals with the dynamics of the environment by learning a transition model that allows for simulation of the environment without interacting with the environment directly. In contrast to model-free approaches, model-based approaches are learned from experience by a function approximation. Theoretically, no specific prior knowledge is required in model-based RL/DRL but incorporating prior knowledge can help faster convergence and better trained model, speed up training time as well as decreasing the required amount of training samples. Also, it is difficult for model-based RL to directly use raw data with pixels as it is high dimensional. This is addressed in DRL by embedding the high-dimensional observations into a lower-dimensional space using autoencoders Finn et al. (2016b). Many DRL approaches have been based on scaling up prior work in RL to high-dimensional problems. A good overview of model-based RL for high-dimensional problems can be found in Plaat et al. (2020), which partitions model-based DRL into three categories: explicit planning on given transitions, explicit planning on learned transitions, and end-to-end learning of both planning and transitions. In general, DRL targets at training DNNs to approximate the optimal policy π^* together with optimal value functions V^* and Q^*. In the following, we will cover the most common model-based DRL approaches including value function and policy search methods.

3.2.1. Value function

We start this category with DQN Mnih et al. (2015) which has been successfully applied to classic Atari and illustrated in Fig.4. DQN uses CNNs to deal with high dimensional state space to approximate the Q-value function.

Monte Carlo tree search (MCTS). MCTS Coulom

(2006) is one of the most popular methods with look-ahead search and it is combined with DNN-based transition model to build a model-based DRL Alaniz (2018). In this work, the learned transition model predicts the next frame and rewards one step ahead using the input of the last four frames of the agent's first-person-view image and the current action. This model is then used by Monte Carlo tree search algorithm to plan the best sequence of actions for the agent to perform.

Value-targeted regression (VTR). Jia et al. (2020) proposed model-based DRL for regret minimization. In their work, a set of models that are 'consistent' with the data collected is constructed at each episode. The consistency is defined as the total squared error, whereas the value function is determined by solving the optimistic planning problem with the constructed set of models.

3.2.2. Policy search

Policy search methods aim to directly find policies by means of gradient-free or gradient-based methods.

Model-ensemble trust-region policy optimization (ME-TRPO). ME-TRPO Kurutach et al. (2018) is mainly based on trust region policy optimization (TRPO) Schulman et al. (2015) which imposes a trust region constraint on the policy to further stabilize learning.

Model-based meta policy optimization (MB-MPO). MB-MPO Clavera et al. (2018) addresses the performance limitation of model-based DRL compared against model-free DRL when learning dynamics models. MB-MPO learns an ensemble of dynamics models and forms a policy that can quickly adapt to any model in the ensemble with one policy gradient step. As a result, the learned policy exhibits less model-bias without the need to behave conservatively.

A summary of both model-based and model-free DRL algorithms is given in Table 2. In this Table, we also categorized DRL techniques into either on-policy or off-policy. In on-policy RL, the policy π^k is updated with data collected by π^k itself. In off-policy RL, each policy has with own data collection, then the data collected from π^0, π^1, ..., π^k is used to trained π^{k+1}.

3.3. Useful techniques to train an agent

In this section, we discuss some useful techniques that are used during training an agent.

Experience replay. Experience replay proposed by Zha et al. (2019) is a useful part of off-policy learning. Experience replay is based on the fact that an agent can learn from certain experiences (transitions , which may be rare but important) more than others (redundant transition or something already learned). By getting rid of as much information as possible from past experiences, it removes the correlations in training data and reduces the oscillation of the learning procedure.

Minibatch learning. Minibatch learning is a common technique that is used together with experience replay. Minibatch allows learning more than one training sample at each step, thus, it helps the learning process robust to outliers and noise.

Target Q-network freezing. As described in Mnih et al. (2015), there are two networks in target Q-network freezing: one network interacts with the environment and another network plays a role of target network. The first network is used to generate target Q-values that are used to calculate losses. The weights of the second target network are fixed and slowly updated with the first network Lillicrap et al. (2015).

Reward clipping. To keep the rewards on a reasonable scale and to ensure proper learning, they are

Table 2. Summary of model-based and model-free DRL algorithms consisting of value-based and policy gradient methods.

DRL Algorithms	Description	Category
DQN Mnih et al. (2015)	Deep Q Network	Value-based, Off-policy
Double DQN van Hasselt et al. (2015)	Double Deep Q Network	Value-based, Off-policy
Dueling DQN Wang et al. (2015)	Dueling Deep Q Network	Value-based, Off-policy
MCTS Alaniz (2018)	Monte Carlo tree search	Valued-based, On-Policy
UCRL-VTRJia et al. (2020)	optimistic planning problem	Valued-based, On-Policy
DDPG Lillicrap et al. (2015)	DQN with Deterministic Policy Gradient	Policy gradient, Off-policy
TRPO Schulman et al. (2015)	Trust Region Policy Optimization	Policy gradient, On-policy
PPO Schulman et al. (2017)	Proximal Policy Optimization	Policy gradient, On-policy
ME-TRPO Kurutach et al. (2018)	Model-Ensemble Trust-Region Policy Optimization	Policy gradient, On-policy
MB-MPO Clavera et al. (2018)	Model-Based Meta- Policy-Optimization	Policy gradient, On-policy
A3C Mnih et al. (2016)	Asynchronous Advantage Actor Critic	Actor Critic, On-Policy
A2C Mnih et al. (2016)	Advantage Actor Critic	Actor Critic, On-Policy

clipped to a specific range (-1 ,1)

4. DRL in Medical Imaging

We start with an exposition of the DRL formulation that is commonly used for parametric medical image analysis tasks such as landmark detection, image registration, and view plane localization. While traditional supervised learning is effective for image detection, segmentation, and classification, formulating these tasks under the DRL framework offers several compelling benefits. First, traditional landmark detection methods perform an exhaustive search over the hypothesis space. As a result, they are inefficient for medical images, which can be several hundred times larger than natural images. In contrast, DRL methods process only a small number of image locations, making them computationally efficient. Second, sequential processing of small image regions drastically reduces the memory footprint. This property will enable scaling up image analysis algorithms to the sizes and resolutions impractical to traditional supervised learning. Third, the DRL formulation can optimally balance time efficiency and accuracy in a principled manner. Applications that require real-time speed or have limited processing power (e.g., edge devices) could significantly benefit from this capability. Finally, DRL's sequential visual search is consistent with the cascade and fixation mechanisms in biological systems, potentially giving rise to more robust features.

DRL also finds its use in other optimization tasks such as hyperparameter tuning, image augmentation selection, neural architecture search, etc., most of which share a common theme of non-differential optimization. Exhaustive grid search for these tasks is time-consuming and DRL is used to learn an efficient search policy. Finally, DRL is used in several miscellaneous topics such as surgical gesture categorization.

Tables 3, 4, and 5 contain a list of 47 selected papers. The selection is made among publications in top journals (such as IEEE Transactions on Medical Imaging and Medical Image Analysis) and conferences (such as MICCAI and CVPR) up to 2020. We also include a few related papers found elsewhere. The list is by no means exhaustive. For each reference, we also provide the task with its concerned image modality and anatomy and offer some remarks when appropriate.

Fig. 7 shows the number of DRL papers published every year, which clearly indicates a growing trend. In most of the listed papers, model-free learning algorithms are used.

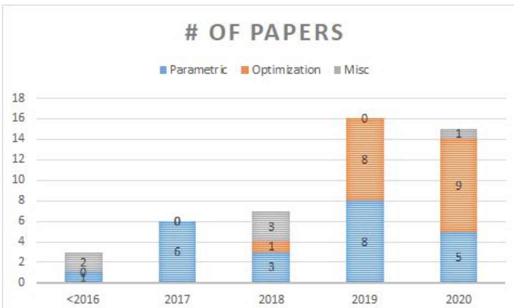

Fig. 7. **The number of DRL papers in medical imaging published. The papers are organized into three categories: parametric medical image analysis ("Parametric"), solving optimization in medical imaging ("Optimization"), and miscellaneous topics ("Misc").**

4.1. DRL for parametric medical image analysis

In many medical image analysis tasks, there are model parameters $\xi = [\xi_1, \xi_2, \ldots, \xi_n]$ to be estimated, given an image I. Table 6 exhibits a collection of common tasks and their associated model parameters. Currently, most model parameters are low dimensional.

Below we first present a general DRL formulation for parametric medical image analysis and then proceed to cover each analysis task in a separate subsection.

4.1.1. Formulation

To formulate a problem into the DRL framework, we have to define three key elements of DRL.

Action. An action $a \in A$, where A is the action space, is what the agent takes to interact with the environment, which is the image I.

One way of defining an action is to move each parameter, say the i^{th} parameter, independently by $\pm\delta\xi_i$ while keeping the other parameters the same. The action space A is given by:

$$A = \{\pm\delta\xi_1, \pm\delta\xi_2, \ldots, \pm\delta\xi_n\}. \tag{20}$$

With this definition, the cardinality of the action space is $|A| = 2n$.

The action space should be specified to guarantee the reachability, that is, starting an initial guess ξ^0, it is possible to reach an arbitrarily-valued parameter, say $\hat{\xi} = [\hat{\xi}_1, \hat{\xi}_2, \ldots, \hat{\xi}_n]$. With the above definition, the reachability is trivially guaranteed, up to quantization error, by taking a series of actions: simply accumulating multiple steps of $\pm\delta\xi_i$ to move the i^{th} parameter by an amount of $\hat{\xi}_i - \xi_i^0$, and repeating this for each of the dimensions.

State. The state is in regard to both the environment and the agent after all actions are taken so far.

Using the action space defined in (20), the agent is at its state ξ_t after taking an action a_t:

$$\xi_t = \xi_{t-1} + a_t = \xi_0 + \sum_{i=1}^{t} a_i. \tag{21}$$

Note that the state of the environment is often chosen as an image (or image patch) 'centered' at ξ_t denoted by $I[\xi_t]$. Other non-centered design choices are possible too.

Reward. In general, the reward function should provide incentive signals when the target is hit or closer and penalize signals otherwise. Designing reward functions for reinforcement learning models is not easy. One design method is called inverse RL Abbeel and Ng (2004) or "apprenticeship learning", which learns a reward function that reproduces observed behaviors.

A commonly used reward function is given as be-

Category I: parameteric medical image analysis			
Task	Reference	Modality	Remarks
Landmark detection	Ghesu et al. (2016, 2017, 2018)	CT & Ultrasound (US)	Multi-scale, using shape constraint
	Alansary et al. (2019)	3D US, fetal head	Multi-scale, multiple DRL strategies
	Vlontzos et al. (2019)	Brain MRI, fetal brain US	Multi-landmark det. via multi-agent RL
	Al and Yun (2019)	CT & MRI	Partial policy-based RL
	Zhang et al. (2020b)	3D ultrasound	landmarking for mitral valve annulus modeling
	Zhang et al. (2020a)	MRI	Fetal pose detection
	Leroy et al. (2020)	MRI and ultrasound	Multi-agent DRL
	Xu et al. (2017)	2D ultrasound	Supervised action classification
Image registration	Liao et al. (2017)	CT and CBCT	Rigid registration
	Krebs et al. (2017)	MR prostate	Nonrigid registration
	Ma et al. (2017)	Depth to CT	2 trans. & 1 rotation parameters
Object/lesion localization & classification	Maicas et al. (2017, 2019)	DCE-MRI	DRL for lesion bounding box detection
	Qaiser and Rajpoot (2019)	WSI microscopy	IHC scoring of HER2
	Xu et al. (2019)	Breast histopathology	RL for selective attention
View plane localization	Alansary et al. (2018)	MR	Hierarchical action steps
	Dou et al. (2019)	Fetal brain US	RL + warm start & active termination
	Huang et al. (2020b)	3D ultrasound	Multi-plane localization
Plaque tracking	Luo et al. (2019)	Intravascular OCT	2 angles with 8 actions
Vessel extraction	Zhang et al. (2018)	CT + MR	Tracing as sequential decision making
	Zhang et al. (2020c)	Coronary CT angiography	Branch-aware Double DQN

Table 3. A summary of references on parameteric medical image analysis using DRL.

Category II: Solving optimization using DRL

Task	Reference	Modality	Remarks
Image and lesion classification	Cheng et al. (2019)	Knee & hip x-ray	Learning to mask an image as classification an attention map
	Pesce et al. (2019)	Chest x-ray	Excluding unlabeled images w/o lesions
	Akrout et al. (2019)	Visual skin image with Q's	RL agent to ask Q's for improved performance
	Ye et al. (2020)	Cervical & lymph node	Synthetic sample selection via RL histopathology
	Wang et al. (2020)	Multimodal ultrasound	Auto-weighting for breast cancer classification
	Ma et al. (2020)	MRI	Longitudinal Alzheimer's disease analysis
Image segmentation	Yang et al. (2019)	CT & MRI, various	Optimizing the DL training strategy segmentation
	Bae et al. (2019)	Brain tumor, heart, prostate	Neural architecture search
	Yang et al. (2020b)	3D ultrasound	DQN-driven catheter segmentation
	Qin et al. (2020)	Kidney tumor segmentation	Automatic data augmentation via DRL
	Liao et al. (2020)	Various	Interactive segmentation with multi-agent RL
Image acquisition	Zaech et al. (2019)	CBCT	Learning to avoid poor images acquisition
	Shen et al. (2018)	CT	Tuning parameters for iterative recon.
	Shen et al. (2020)	CT	Learning to scan
	Pineda et al. (2020)	MRI	Active k-space sampling
	Li et al. (2020)	MRI	Pixel-wise operations using RL
Radiotherapy planning	Shen et al. (2019)	n/a	Tuning parameters for inverse treatment planning
Video summarization	Liu et al. (2020a)	Ultrasound video	Summarization using DRL

Table 4. A summary of references on solving optimization using DRL.

Category III: Miscellaneous topics

Task	Reference	Modality	Remarks
Surgical gesture	Liu and Jiang (2018)	Surgical video	Small/large time step
Personalized	Zhu et al. (2018)	Smart device data	Group-driven RLmHealth
Model	Neumann et al. (2015, 2016)	n/a	Heart modeling personalization
	Abdi et al. (2018)	n/a	Muscle excitation estimation
	Joos et al. (2020)	n/a	Musculoskeletal control

Table 5. A summary of references of miscellaneous topics that utilize DRL.

Task	Parameters
2D landmark detection	$\xi = [x, y]$
3D landmark detection	$\xi = [x, y, z]$
Rigid 2D object detection	$\xi = [x, y, \alpha, s]$
Rigid 3D object detection	$\xi = [x, y, z, \alpha, \beta, \gamma, s]$
Rigid 2D/3D registration	$\xi = [x, y, z, \alpha, \beta, \gamma]$
View plane localization in 3D	$\xi = [a, b, c, d]$
Others	ξ depends on the task

Table 6. Common medical image analysis tasks and their associated model parameters. x, y, z **are for translation,** α, β, γ **for rotation,** s **for scaling, and** $[a, b, c, d]$ **for depicting a plane.**

low:

$$R(s_t, s_{t-1}, a_t) = D(\xi_{t-1}, \hat{\xi}) - D(\xi_t, \hat{\xi}), \tag{22}$$

where $D(\xi_1, \xi_2)$ is a distance function that measures the difference between ξ_1 and ξ_2. If certain action reduces the difference, then a positive reward is obtained; otherwise, a negative reward is obtained.

To further intensify the effect of reward especially when the change in the difference is small, one can use

$$R'(s_t, s_{t-1}, a_t) = sgn(R(s_t, s_{t-1}, a_t)), \tag{23}$$

where $sgn(x)$ takes the sign of the value x. So, if certain action reduces the difference, then a positive reward $+1$ is obtained; otherwise, a negative reward -1 is obtained.

Once we have these three elements, we can invoke the DQL algorithm to trigger the learning process. Once the Q-function is learned, we can choose the action that maximizes the Q-function at each iteration.

It is clear that the search trajectory (or path) is implicitly related to the three elements. An alternative is to make the path explicit, that is, path supervision Liao et al. (2017); Xu et al. (2017). One path supervision approach is to guide the selection of the action that maximizes the reward in a greedy fashion for every iteration.

$$\begin{aligned} \hat{a}_t &= \arg\max_a R(s_t, s_{t-1}, a_t) \\ &= \arg\max_a D(\xi_{t-1}, \hat{\xi}) - D(\xi_{t-1} + a, \hat{\xi}). \end{aligned} \tag{24}$$

This converts a reinforcement learning problem into supervised learning. With the pairs of $(I[\xi_{t-1}], \hat{a}_t)$ forming training data, we can train supervised classification or regression functions.

4.1.2. Landmark detection

Medical landmarks are commonly used to represent distinct points in an image that likely coincide with anatomical structures. In clinical practices, landmarks play important roles in interpreting and navigating the image just like geographic landmarks that help travelers navigate the world. Also, landmarks are used to derive measurements (e.g., width, length, size, etc.) of organs Xu et al. (2017), and to trigger subsequent, computationally intensive medical image analysis applications. In multi-modality image registration (such as PET-CT) or in registration of follow-up scans, the fusion of multiple images can be initialized or guided by the positions of such anatomical structures Johnson and Christensen (2002); Crum et al. (2004). In vessel centerline tracing, detected vessel bifurcations Liu et al. (2010) can provide the start and end points of certain vessels to enable fully-automated tracing Beck et al. (2010). In organ segmentation, the center position of an organ can provide the initial seed points to initiate segmentation algorithms Banik et al. (2009). Landmark points situated on the organ surface, once detected, offer better initialization for segmentation Lay et al. (2013). In seminar reporting, automatically found anatomical structures can be helpful in configuring the optimal intensity window for display Pauly et al. (2011); Lay et al. (2013), or offer the text tooltips for structures in the scan Seifert et al. (2010).

Artificial agent. In a series of papers, Ghesu et al. Ghesu et al. (2016, 2017, 2018) present a multi-scale approach for detecting anatomical landmarks in a 3D volume using an artificial agent. The landmark is represented as a 3D point and the actions include moving one-voxel step to the left, right, up, down, and forward and back. The reward function is given by (22).

At each scale, a scale-specific Q-function is learned to enable the agent to effectively search for objects in the image, as opposed to scanning the volumetric space exhaustively. Per scale-space theory, the system captures global context on coarse scale and local context on fine scale. The search starts at the coarsest scale level, where the search model is trained for convergence from any starting point in the image. On this scale level the field of view of the agent is very large with sufficient global context to ensure effective navigation. Upon convergence, the scale level is changed to the next level and the search continues. The process is repeated on the following scales until convergence on the finest scale.

The convergence criterion is met when trajectories converge on small, oscillatory-like cycles. Once such a cycle is identified at detection time, the search is stopped and the location is recorded as the detection result. An interesting finding is that, when searching for a landmark outside of the present scan, the search trajectory leaves the image space, signaling that the landmark is missing from the field-of-view. To guarantee this consistent behavior, the system is trained by differently cropped images.

In addition, the constrained spatial distribution of anatomical landmarks using statistical shape model-

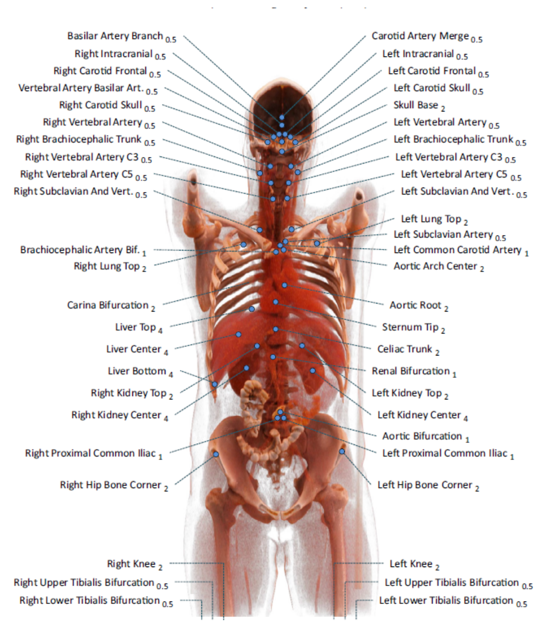

Fig. 8. The list of 49 anatomical landmarks. Courtesy of Ghesu et al. (2018).

ing and robust estimation theory Torr and Zisserman (2000) is used to offer a probabilistic guarantee on the spatial coherence of the identified landmarks and to recognize if there are landmarks missing from the field-of-view. This shape fitting further makes the detection of landmarks more robust.

The proposed method is tested on detecting a cohort of 49 landmarks (see Figure 8) in a complete dataset of 5,043 3D-CT scans over 2,000 patients. When evaluating the detection performance, the landmarks 3cm within the image border are ignored. This value was selected in agreement with our expert annotators. Perfect detection results with no false positives or negatives are reported. Figure 9 shows the detection results of two vascular landmarks on three different levels from left to right, which demonstrates the preciseness of the approach.

Alansary et al. (2019) evaluate different reinforcement learning agents with different training strategies for detecting anatomical landmarks in 3D images. The

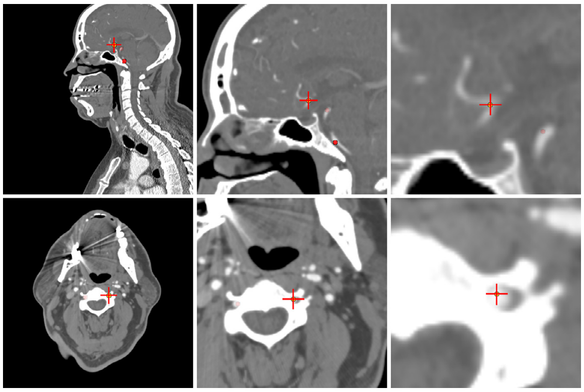

Fig. 9. Visualization of detection results of two vascular landmarks on 3 different levels from left to right. Courtesy of Ghesu et al. (2018).

specific training strategies include DQN, DDQN (Double DQN), Duel DQN, and Duel DDQN. Also fixed- and multi-scale optimal path search strategies are compared. The finding is that the optimal DQN architecture for achieving the best performance depends on the environment.

Vlontzos et al. (2019) consider the interdependence between multiple landmarks as they are associated with the human anatomy. It is likely that localizing one landmark helps detect the other landmarks. They propose to train a set of multiple collaborative agents using reinforcement learning in order to detect multiple landmarks, instead of a naive approach that learns many separate agents, one agent for each landmark. It is shown that the multi-agent RL achieves significantly better accuracy by reducing the detection error by 50% on detecting 7-14 landmarks for three tasks, consumes fewer computational resources, and reduces the training time, when compared with the naive approach.

In Al and Yun (2019), an RL agent is learned for landmark localization in 3D medical images, following the formulation in Ghesu et al. (2017). However, an actor-critic approach is utilized to directly approximate the policy function In addition, in order to speed up the learning and reach a more robust localization, multiple partial policies on different sub-action spaces are learned instead of a single complex policy on the original action space. For a 3D landmark (x, y, z), the action space is $A = \{\pm\delta x, \pm\delta y, \pm\delta z\}$; so it is natural to define three sub-action spaces $A_k = \{\pm\delta k\}$ with $k \in \{x, y, z\}$ by projecting the actual action space onto different Cartesian axes. Experiments on three datasets, namely 71 contrast-enhanced coronary CT angiography volumes with 8 landmarks, 150 cardiac CT volumes with a landmark of left atrial appendage (LAA) seed-point, and 18 MR spine images with 5 lumbar vertebra landmarks, demonstrate that the proposed actor-critic approach with partial policies achieves robust and improved performances, compared to the conventional actor-critic and widely used deep Q-learning approach.

Zhang et al. (2020b) propose a bottom-up approach for automatically building a mitral valve annulus model from 3D echocardiographic images in real time, in which the very first step is to automatically detect a few key landmarks associated with the above annulus model using the artificial agent Ghesu et al. (2017).

Zhang et al. (2020a) incorporate priors on physical structure of the fetal body to optimize multi-agent for detection of fetal landmarks. In this work, they use graph communication layers to improve the communication among agents based on a graph where each node represents a fetal body landmark. The proposed network architecture contains two parts corresponding to shared CNNs for feature extraction and graph communication networks to merge the information of correlated landmarks. Furthermore, the distance between agents and physical structures such as the fetal limbs is used as a reward. The evaluation is conduction on 19,816 3D BOLD MRI volumes acquired on a 3T Skyra scanner. The proposed method achieves an average detection accuracy of 87.3% under a 10-mm threshold and 6.9mm as the mean error. In Leroy et al. (2020), a communicative multi-agent reinforcement learning method is proposed for detecting landmarks in an adult MRI and fetal ultrasound brain image. The experiments demonstrate the use of multiple cooperating agents by learning their communication with each other outperforms previous approaches that are based on single agents.

Supervised action classification. Xu et al. (2017) propose to approach landmark detection as image partitioning. This nontrivial approach is derived from path supervision.

Consider an agent that seeks an optimal action path from any location at (x, y) towards a landmark $l = (\hat{x}, \hat{y})$, which is composed of optimal action steps at pixels along the path on an image grid Ω. In other words, at each pixel the agent is allowed to take an action a with a unit movement $d_x^{(a)} \in \{-1, 0, 1\}$ and $d_y^{(a)} \in \{-1, 0, 1\}$. With the constraint of $\|d_x^{(a)}\|^2 + \|d_y^{(a)}\|^2 = 1$, we basically allow four possible action types $a \in \{0, 1, 2, 3\}$:

$$\begin{aligned} UP: &\quad (d_x^{(0)} = 0,\ d_y^{(0)} = -1), \\ RIGHT: &\quad (d_x^{(1)} = 1,\ d_y^{(1)} = 0), \\ DOWN: &\quad (d_x^{(2)} = 0,\ d_y^{(2)} = 1), \\ LEFT: &\quad (d_x^{(3)} = -1,\ d_y^{(3)} = 0). \end{aligned}$$

The optimal action step $\hat{a}$ is selected as the one with minimal Euclidean distance to the landmark l after its associated movement,

$$\hat{a} = \arg\min_a \sqrt{(x - \hat{x} + d_x^{(a)})^2 + (y - \hat{y} + d_y^{(a)})^2}. \quad (25)$$

Simple derivations show that the selection of $\hat{a}$ falls

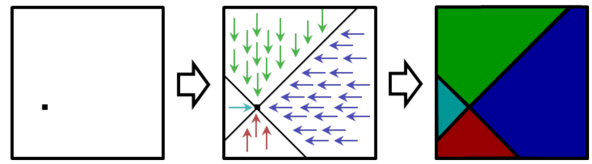

Fig. 10. The discrete action map representation.

into four regions (one for each action type), where the regions are partitioned by two lines with slopes of ± 1 crossing the landmark (Figure 10):

$$y = x + (\hat{y} - \hat{x}), \quad y = -x + (\hat{x} + \hat{y}).$$

This generates a discrete action map $a(x, y)$ that represents the pixel-wise optimal action step moving toward the target landmark location.

During training to estimate the action map for a given image, a fully convolutional neural network, called a deep image-to-image network (DI2IN), can be employed given its efficient sampling scheme and large receptive field for comprehensive feature learning. During testing, the landmark location needs to be derived from the estimated action map. To this, an aggregate approach is proposed. With the output action map $A(x, y)$ from DI2IN, the estimated landmark location coordinates (x', y') are determined by maximizing an objective function $C(\cdot)$ summed up with that of each action type $C_a(\cdot)$.

$$(x', y') = \arg\max_{(x,y)} C(x, y) = \arg\max_{(x,y)} \sum_a C_a(x, y), \quad (26)$$

where the action-wise objective function at pixel (x, y) is aggregated by the pixels with that specific action on the same row or column, specifically

$$C_a(x, y) = \begin{cases} d_x^{(a)}\{\sum_i (2\,\pi[i \geq x] - 1)\pi[A(i, y) == a]\} \\ \qquad \text{if } \|d_x^{(a)}\| = 1, \\ d_y^{(a)}\{\sum_j (2\,\pi[j \geq y] - 1)\pi[A(x, j) == a]\} \\ \qquad \text{if } \|d_y^{(a)}\| = 1, \end{cases} \quad (27)$$

where $\pi[.]$ is an indicator function. Such aggregation enables robust location coordinate derivation even with a suboptimal action map from the DI2IN output.

In experiments on detecting landmarks from a cardiac or obstetric ultrasound image in two datasets with 1,353 and 1,642 patients, respectively, it is demonstrated that the proposed approach achieves the best results when compared with then state-of-the-art approaches that include the artificial agent.

4.1.3. *Image registration*

Robust image registration in medical imaging is essential for the comparison or fusion of images, acquired from various perspectives, in different modalities or at different times. In terms of modeling the registration, there are two ways: rigid and non-rigid.

Rigid registration. Rigid registration is fully specified by a few number of transformation parameters. For example, a 3D rigid registration typically has 6 parameters to optimize. Traditionally, image registration is solved by optimizing an image matching metric such as normalized correlation coefficient or mutual information as a cost function, which is difficult due to the non-convex nature of the matching problem.

Liao et al. (2017) propose an artificial agent to perform image registration. It casts the image registration problem as a process of finding the best sequence of motion actions (*e.g.,* up, down, left, right, etc.) that yields the desired image registration parameter. The input to the agent is the 3D raw image data and the current estimate of image registration parameter, and the output of the agent, which is modeled using a deep convolutional neural network, is the next optimal action. Further, it utilizes the path supervision approach to supervise the end-to-end training. Since the agent is learned, it avoids the issue of current approaches that are often customized to a specific problem and sensitive to image quality and artifacts.

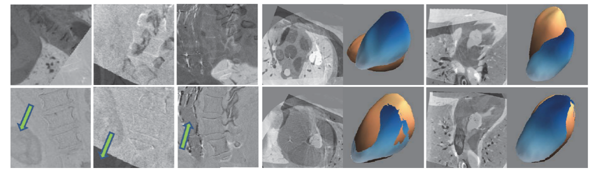

Fig. 11. Registration examples shown as the difference between the reference and floating images, before (upper row) and after (lower row) registration. The mesh overlay before and after registration is shown for cardiac use case for improved visualization. Picture courtesy of Liao et al. (2017).

In experiments, the proposed approach is evaluated on two datasets: spine (87 pairs of images) and heart (97 pairs of images). In the first dataset of aligning abdominal spine CT and CBCT, the main challenging lies in that CT has a much larger FOV than CBCT, leading to many local optima in the registration space due to the repetitive nature of the spine. In the second dataset of registering cardiac CT and CBCT (as in Figure 11), the main challenge lies in the poor quality of CBCT with severe streaking artifacts and weak soft tissue contrast at the boundary of the epicardium. On both datasets, the artificial agent outperforms several state-of-art registration methods by a large margin in terms of both accuracy and robustness.

Similarly, Ma et al. (2017) use the artificial agent to register a 2.5D depth image and a 3D CT. Different from Liao et al. (2017), it uses dueling DQN to learn the Q function instead of path supervision. Further, although it involves a six-degree-of-freedom transformation, the search space is simplified into two translations and one rotation as the rest of the transformation can be determined/inferred through the sensor calibration process together with the depth sensor readings. It also invokes orthographic projection to generate 2D images that are fed into the Q function. Quantitative evaluations are conducted on 1788 pairs of CT and depth images from real clinical setting, with 800 as training. The proposed method achieves state-of-the-art performance, when compared with several approaches including Ghesu et al. (2016).

Non-rigid registration. When a rigid transformation is insufficient to describe the transformation between two images, a non-rigid registration comes into play, which has more than 6 parameters in 3D to optimize, depending on the class of non-rigid registration.

Krebs et al. (2017) extend the artificial agent approach to handle non-rigid registration. In particular, the parametric space of a statistical deformation model for an organ-centered registration of MR prostate images is explored. There are $m = 15$ PCA modes in 2-D and $m = 25$ modes in 3-D kept to model the prostate deformation, with $2 \times m$ actions are defined.

To tackle the difficulty of obtaining trustworthy ground-truth deformation fields, Krebs et al. (2017) proceed with a large number of synthetically deformed image pairs derived from only a small number of inter-subject pairs. Note that the extracted ground truth reaches a median DICE coefficients of 0.96 in 2-D and 0.88 in 3-D. The Q function is then learned.

The algorithm is tested on inter-subject registration of prostate MR data (41 3D volume in total with 8 for testing, resulting in 56 inter-subject pairs). For the 2D experiment, the middle slice of each volume is utilized. Before the non-rigid registration, the initial translation registration is performed using the Elastix approach Klein et al. (2010) by registering each of the test images to an arbitrarily chosen template from the training database. The final registration result reaches a median DICE score of 0.88 in 2-D and 0.76 in 3-D, both better than competing state-of-the-art registration algorithms.

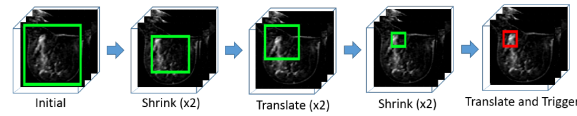

Fig. 12. The illustration of the detection process, with the learnt DRL agent outputting a series of allowable actions to realize final detection of a 3D lesion. Picture courtesy of Maicas et al. (2017).

4.1.4. *Object/lesion localization and detection*

DRL is also leveraged to detect objects Jie et al. (2016). Maicas et al. (2017) presented such an approach for detecting breast lesions from dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI).

The bounding box for a 3D lesion is defined as $\mathbf{b} = [b_x, b_y, b_z, b_w, b_h, b_d]$, where b_x, b_y, b_z denote the top-left-front corner and b_w, b_h, b_d define the lower-right-back corner of the bounding box. The actions are defined as $\{l_x^+, l_x^-, l_y^+, l_y^-, l_z^+, l_z^-, s^+, s^-, w\}$, where l, s, w are translation, scale and trigger actions, with the subscripts x, y, z denoting the horizontal, vertical or depth translation, and superscripts $+, -$ meaning positive or negative translation and up or down scaling. The signed reward function is used. DQN is learned based on the ResNet architecture.

Experiments are conducted on DCE-MRI volumes from 117 patients. The training set contains 58 patients annotated with 72 lesions, and the testing set has 59 patients and 69 lesions. Results show a similar accuracy to state-of-the-art approaches, but with significantly reduced detection time.

Pesce et al. (2019) study how to localize pulmonary lesions in a chest radiograph. In one of the proposed methods, a recurrent attention model with annotation feedback (RAMAF) is learned using RL to observe a short sequence of image patches. The classification score is used as a reward signal, which penalizes the exploration of areas that are unlikely to contain nodules and encourages the learning of a policy that maximizes the conditional probability of the true label given a series of image patches within the radiographs.

In Qaiser and Rajpoot (2019), a sequential learning task is formulated to estimate from a giga-pixel whole slide image (WSI) the immunohistochemical (IHC) scoring of human epidermal growth factor receptor 2 (HER2) on invasive breast cancer (BC), which is a significant predictive and prognostic marker. To solve this task, DRL is employed to learn a parameterized policy to identify diagnostically relevant regions of interest (ROIs) based on current inputs, which are comprised of two image patches cropped at 40× and 20× magnification levels. The selected ROIs are processed by a CNN for HER2 scores. This avoids the need to process all the sub-image patches of a given tile and saves a large of amount of computations. Refer to Figure 13 for some illustrative results of HER2 scoring.

Xu et al. (2019) take the computational challenge of breast cancer classification from a histopathological image. Due to the large size of a histopathological image, pathologists in clinical diagnosis first find an abnormal region and then investigate the detail within the region. Such a human attention mechanism inspires an attention-based deep learning approach. It consists of two networks for selection and classification tasks separately. The selection network is trained using DRL, which outputs a soft decision about whether the cropped patch is necessary for classification. These selected patches are used to train the classification network, which in turn provides feedback to the selection network to update its selection policy. Such a co-evolution training strategy enables fast convergence and high classification accuracy. Evaluation based on a

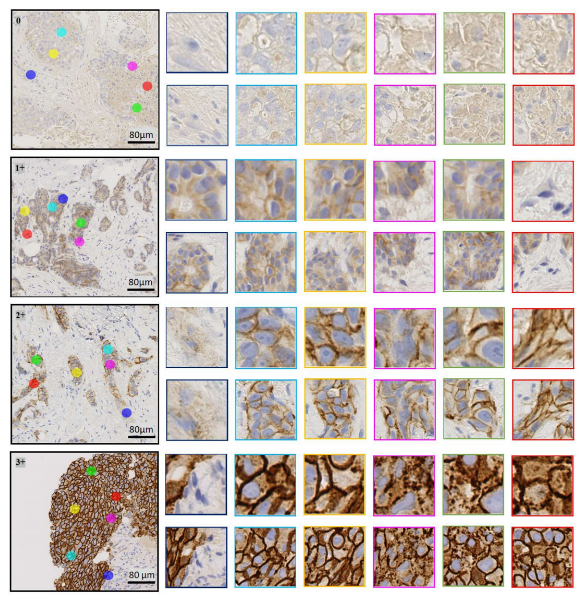

Fig. 13. Example of four image tiles with selected regions-of-interest (ROIs) predicted by Qaiser and Rajpoot (2019), for each HER2 score (0-3+), respectively. The first column shows the input images and colored disks show the predicted locations. The remaining columns show the selected regions at 40× and 20× around the selected locations. The first selected region is shown with blue bounding boxes and the last selected region is shown with red bounding boxes. Picture courtesy of Qaiser and Rajpoot (2019).

public breast cancer histopathological image database of 7,909 images and eight subclasses of breast cancers from 82 patients (58 malignant and 24 benign) demonstrates about 98% classification accuracy while only taking 50% of the training time of the previous hard-attention approach.

4.1.5. View plane localization

Alansary et al. (2018) propose to use DRL to detect canonical view planes in MR brain and cardiac volumes. A plane in 3D $ax + by + cz + d = 0$ is parameterized by a 4D vector $[a, b, c, d]$. The eight actions are defined as $\{\pm\delta_{\xi_x}, \pm\delta_{\xi_y}, \pm\delta_{\xi_z}, \delta_d, \}$, which update the plane parameters as $a = cos(\xi_x + \delta_{\xi_x})$, $b = cos(\xi_y + \delta_{\xi_y})$, $c = cos(\xi_z + \delta_{\xi_z})$, and $d = d + \delta_d$. The signed reward function is used. Further a multi-scale strategy is utilized, with the action steps are refined a coarse-to-fine fashion.

The experiments are based on 382 brain MR volumes (isotropic 1mm) and 455 short-axis cardiac MR volumes ($1.25 \times 1.25 \times 2mm^3$). Figure 14 visualize the viewing planes to be detected. The specific Q-learning strategies include DQN, DDQN (Double DQN), Duel DQN, and Duel DDQN. The detection of the anterior-posterior commissure (ACPC) and mid-sagittal planes reach an error less than 2mm and the detection of the apical four chamber plane reaches an error around 5mm, where the error is measured as the distance between anatomical landmarks and the detected planes and the landmarks are accordingly specified for the ACPC and mid-sagittal planes.

Dou et al. (2019) study how to use a DRL agent to localize two standard planes of transthalamic (TT) and transcerebellar (TC) positions in a 3D ultrasound volume of the fetal head. The plane parameterization, action space, and reward function are defined in a similar manner to Alansary et al. (2018). To ease the localization, they propose to augment the agent with a warm start module for better initialization and an active termination module for drift prevention. Based on their extensive validation on in-house datasets of 430 prenatal US volumes, the proposed approach improves both the accuracy and efficiency of the localization system.

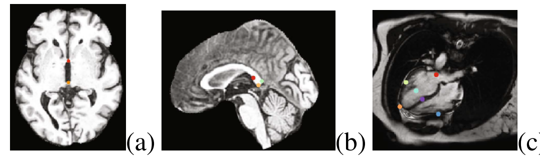

Fig. 14. The viewing planes for detection: (a) Brain axial ACPC plane, (b) Brain mid saggital plane (c) Cardiac apical four chamber plane. The landmarks are visualized for better definition of the plane and used for error calculation. Picture courtesy of Alansary et al. (2018).

Huang et al. (2020a) localize multiple uterine standard planes in 3D ultrasound simultaneously by a multi-agent DRL, which is equipped by one-shot neural architecture search (NAS) module. In this work, gradient-based search using a differentiable architecture sampler (GDAS) is employed to accelerate and stabilize the training process. Furthermore, to improve the system robustness against the noisy environment, a landmark-aware alignment model is utilized. The spatial relationship among standard planes is learned by a recurrent neural network (RNN). They conduct the experiment on an in-house dataset of 683 volumes which show that multiple agents with recurrent network obtain the best performance.

4.1.6. Plaque tracking

Analysis of atherosclerotic plaque in clinical application relies on the use of Intravascular Optical Coherence Tomography (IVOCT), in which a continuous and accurate plaque tracking algorithm is necessary. However, it is challenging to do so due to speckle noise, complex and various intravascular morphology, and a large number of IVOCT images in a pullback. The detected plaque section is represented as a sector with unified radius and the sector is represented as two-tuples $d = (\Theta_S, \Theta)$, where Θ denotes the scale (included angle) of the detected sector, $\Theta_S \in [0, 2\pi]$ denotes the localization (starting angle on the polar coordinate space) of the detected sector. The eight transform actions are Bidirectional Expansion (BE), Bidirectional Contraction (BC), Contra Rotation (COR), Clockwise Rotation (CLR), Contra Unilateral Expansion (COUE), Clockwise Unilateral Expansion (CLUE), Clockwise Unilateral Contraction (CLUC), and Contra Unilateral Contraction (COUC). The reward function is defined as

$$R = \begin{cases} 1 & if\ IOU(d^a, g) - IOU(d, g) > 0; \\ -1 & if\ IOU(d^a, g) - IOU(d, g) < 0; \\ 1 & if\ IOU(d^a, g) - IOU(d, g) = 0 \\ & \&\ IOU(d^a, g) > 0.95; \\ -1 & if\ IOU(d^a, g) - IOU(d, g) = 0 \\ & \&\ IOU(d^a, g) < 0.95, \end{cases} \tag{28}$$

where g is the ground truth sector region, d is the current detected sector, and d^a is the next detected sector based on current selected action. $IOU(d^a, g)) = IOU(d, g)$ only happens when stop action is selected. Fig. 15 is the proposed DRL framework.

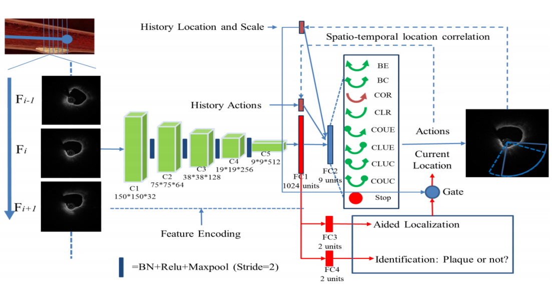

Fig. 15. The DRL framework is proposed to leverage the spatiotemporal information to achieve continuous and accurate plaque tracking. Picture courtesy of Luo et al. (2019).

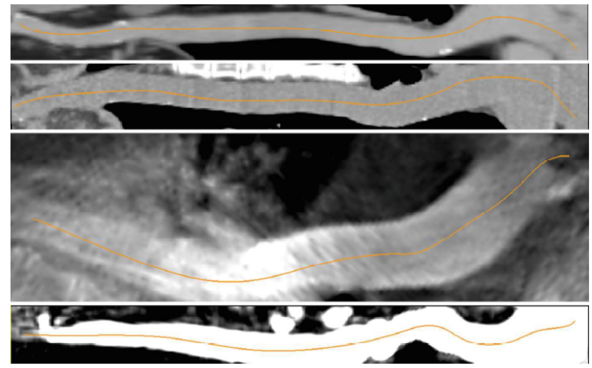

Fig. 16. Example of traced aorta centerlines in the curved planar reformatting (CPR) view. Picture courtesy of Zhang et al. (2018).

4.1.7. Vessel centerline extraction

Zhang et al. (2018) propose to use deep reinforcement learning for vessel centerline tracing in multi-modality 3D volumes. The ground truth vessel center points are given as $\mathbf{G} = [\mathbf{g_0}, \mathbf{g_1}, \ldots, \mathbf{g_n}]$. The key idea is to learn a navigation model for an agent to trace the vessel centerline through an optimal trajectory $\mathbf{P} = [\mathbf{p_0}, \mathbf{p_1}, \ldots, \mathbf{p_m}]$. The action space is defined as $\mathcal{A} = \{left, right, top, bottom, front, back\}$, that is, moving to one of six neighboring voxels.

For the current point $\mathbf{p_t}$, a corresponding point $\mathbf{g_d}$ on the centerline that has the minimum distance to the point $\mathbf{p_t}$ is first found. A point-to-curve measure is then defined as

$$D(\mathbf{p_t}, \mathbf{G}) = \|\lambda(\mathbf{p_t} - \mathbf{g_{d+1}}) + (1 - \lambda)(\mathbf{g_{d+2}} - \mathbf{g_d})\|. \tag{29}$$

It consists of two terms. The first term pulls the agent position towards the ground truth centerline and the second term enforces the agent towards the direction of the curve. With the aid of $D(\mathbf{p_t}, \mathbf{G})$, the reward function is given as

$$r_t = \begin{cases} D(\mathbf{p_t}, \mathbf{G}) - D(\mathbf{p_{t+1}}, \mathbf{G}), & if\ \|\mathbf{p_t} - \mathbf{g_d}\| \le l \\ \|\mathbf{p_t} - \mathbf{g_d}\| - \|\mathbf{p_{t+1}} - \mathbf{g_d}\|, & otherwise \end{cases} \tag{30}$$

For evaluation, the authors collect 531 contrasted

CT, 887 non-contrasted CT, 737 C-arm CT, and 232 MR volumes from multiple sites over the world. For the original 12-bit images, the voxel intensity is clipped and normalized within [500,2000]. The intensity distribution of MR is mapped to that of CT. All these volumes are then mixed for training and testing. The proposed algorithm achieves better performance when compared with a supervised 3D CNN approach.

Recently, Zhang et al. (2020d) make use of DDQN and 3D dilated CNN to address the problem of accurate coronary artery centerline. Their network consists of two parts: a DDQN-based tracker to predict the next action and a branch-aware detector to detect the branch points and radius of coronary artery. With such network architecture, it requires only one seed as input to extract an entire coronary tree. The two-branch network has been evaluated on CAT08 challenge and obtains state-of-the-art performance while it costs only 7s for inference. Fig. 16 shows an example of traced aorta centerlines in the curved planar reformatting (CPR) view.

4.2. Solving optimization using DRL

Because DRL can handle the non-differential metrics, it is widely used to solve optimization problems where conventional methods fall apart. Table 4 is an array of such applications including tuning hyperparameters for radiotherapy planning, selecting the right image augmentation selection for image classification, searching best neural architecture for segmentation, and avoiding poor images via a learned acquisition strategy.

4.2.1. Image classification

Akrout et al. (2019) propose to integrate a CNN classification model with a RL-based Question Answering (QA) agent for skin disease classification. To better identify the underlying condition, the DNN-based agent learns how to ask the patient about the presence of symptoms, using the visual information provided by CNN and the answers to the asked questions. It is demonstrated that the integrated approach increases the classification accuracy over 20% when compared to the CNN-only approach that uses only the visual information. It narrows down the diagnosis faster in terms of the average number of asked questions, when compared with a conventional decision-tree-based QA agent.

Cheng et al. (2019) study how to use semantic segmentation that produces a hard attention map for improved classification performance. In particular, a segmentation agent and a classification model are jointly learned. The segmentation agent, which produces a segmentation mask, is trained via a reinforcement learning framework, with reward being the classification accuracy. The classification model is learned using both original and masked data as inputs. Promising results are obtained on Stanford MURA dataset, consisting of 14,863 musculoskeletal studies of elbows, finger, forearm, hand, humerus, shoulder, and wrist with 9,045 normal and 5,818 abnormal labeled cases and on a hip fracture dataset, consisting of 1,118 pelvic radiographs with 6 classes: no fracture, intertrochanteric fracture, displaced femoral neck fracture, non-displaced femoral neck fracture, arthroplasty, and ORIF (previous internal fixation). Fig. 17 shows some sample X-Ray images and their corresponding attention maps.

To combat the issue of data shortage in medical image classification, synthesizing realistic medical images offers a viable solution. Ye et al. (2020) investigate the issue of synthetic sample selection for im-

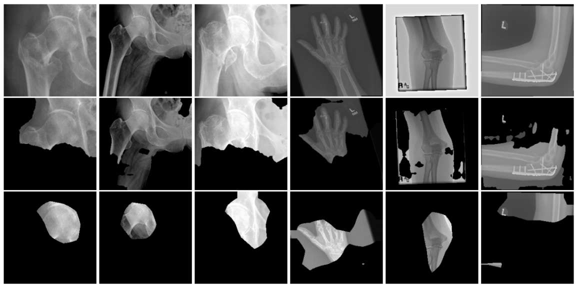

Fig. 17. X-Ray examples (top) and the masks created by Cheng et al. (2019) (middle) and DenseNet+GradCam (bottom) for hip, hand, and elbow. Picture courtesy of Cheng et al. (2019).

proved image classification in order to assure the quality of synthetic images for data augmentation purposes because some of the generated images are not realistic and pollute the data distribution. The authors train a DRL agent via proximal policy optimization (PPO) to choose synthetic images containing reliable and informative features, using the classification accuracy as the reward. Extensive experiments are conducted on two image datasets of cervical and lymph node histopathology images and the performances are improved by 8.1% and 2.3%, respectively.

Wang et al. (2020) combines four different types of ultrasonography to discriminate between benign and malignant breast nodules by proposing a multi-modal network. In their network, the modalities interact through a RL framework under weight-sharing, i.e., automatically find the optimal weighting across modalities to increase accuracy. Corresponding to four modalities, there are four streams (ResNet18 is used as backbone) and each stream provides one loss. Together with four losses from four streams, there is another fusion loss. All the five losses are weighted by coefficients which are automatically learned through an RL framework. The auto-weighting network is evaluated on 1,616 sets of multi-modal ultrasound images of breast nodules and it shows that multi-modal methods outperform single-modal methods.

4.2.2. *Image segmentation*

Medical image segmentation aims at finding the exact boundary of an anatomical or pathological structure in a medical image. In the most general form, an image segmentation approach assigns semantic labels to pixels. By grouping the pixels with the same label, object segmentation is realized. From image segmentation, clinical measurements such as organ volume can be computed and diseases such as enlarged liver can be diagnosed.

There are early approaches that use RL for image segmentation Shokri and Tizhoosh (2003); Sahba et al. (2006); Wang et al. (2013), based on a limited number of parameters to derive image segmentation results. This severely limits the segmentation performances. Contemporary medical image segmentation methods are based on machine learning Zhou (2010) or fully convolutional deep network structures such as U-Net Ronneberger et al. (2015). However, there are a few strategic choices to make in U-Net training, such as tuning the learning rate, data augmentations, data pre-processing, etc. Previous methods are based either on extensive experimentation and grid parameter search or heuristics stemming from specific domain knowledge and expertise; Yang et al. (2019) present a RL searching approach to optimize the training strategy for 3D medical image segmentation, which boosts the performance of the baseline models.

Neural architecture search (NAS) Zoph and Le (2016) automates the task of designing neural networks for a special application, often leading to better performance. However, NAS is seldom applied to medical image segmentation. Bae et al. (2019) make such an attempt, aiming to modify a U-Net base ar-

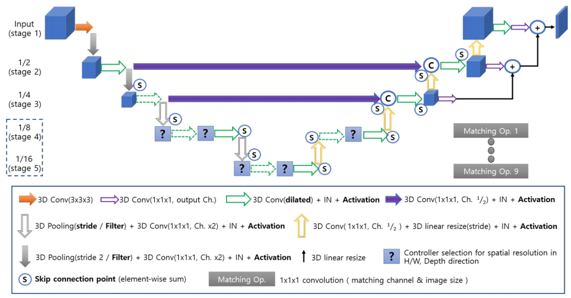

Fig. 18. The proposed base architecture that is modified to best fit the application by using RL. Picture courtesy of Bae et al. (2019).

chitecture as in Fig. 18 so that the image segmentation performance is improved. The search space constitutes multiple factors, including input size, pooling type, filter size, and stride size, activation type, skip connection point, and dilation rate. Using the searched U-Net, the segmentation performances on the medical segmentation decathlon (MSD) challenges are better than those of the nnU-Net approach Isensee et al. (2018), which is considered as state-of-the-art approach.

The lack of labeled data is one of the biggest challenges in medical image segmentation. Among existing methods that intend to increase and diversify the available training samples, augmentation has been commonly used Yang et al. (2019); Ravishankar et al. (2017). However, data augmentation has been applied as pre-processing and there is no guaranteed that it is optimal. In order to learn an optimal augmentation under an end-to-end segmentation framework, Qin et al. (2020) propose to train both augmentation and segmentation modules simultaneously and use the errors in segmentation procedure as feedback to adjust the augmentation module. In addition to scarce annotation, class-imbalance issue is also addressed in Dual-Unet Yang et al. (2020a), which proposes a semi-supervised approach that leverages RL as a pre-localization step for catheter segmentation. Dual-Unet is trained on both limited labeled and abundant unlabeled images with a two-stage procedure.

By iteratively incorporating user hints, Liao et al. (2020) propose IteR-MRL with multi-agent reinforcement learning to capture the dependency among voxels for segmentation task as well as to reduce the exploration space to a tractable size.

4.2.3. Image acquisition and reconstruction

CT metal artifacts, whose presence affects clinical decision making, are produced because of there is an inconsistency between the imaging physics and idealized assumption used in CT reconstruction algorithm. While there are many metal artifact reduction (MAR) algorithms in the literature that post-process the already acquired data say from a pre-determined cone beam CT imaging trajectory or reconstructed images, Zaech et al. (2019) propose to design a task-aware, patient-specific imaging trajectory in order to avoid acquiring "poor" images that give rise to beam hardening, photon starvation, and noise. Such a design strategy is learned offline via a DRL agent that predicts the next acquisition angle that maximizes a final detectability score. Fig. 19 compares the reconstructed images from a straightforward short-scan and a task-aware trajectory recommended by the agent. It is clear that the metal artifacts are reduced.

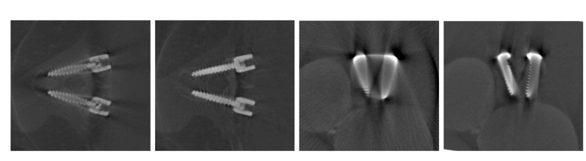

Fig. 19. Two examples of axial slices from a volume reconstructed from (a,c) a straightforward short-scan and (b,d) a task-aware trajectory recommended by the agent. It is evident that the visual quality of the images reconstructed by using the agent is better. Picture courtesy of Zaech et al. (2019).

CT iterative reconstruction solves an optimization problem that uses a total variation (TV) regularization Rudin et al. (1992):

$$f^* = \arg\min_f \frac{1}{2}|Pf - g|^2 + |\lambda \cdot \nabla f|, \tag{31}$$

where f^* is the image to be reconstructed, P is the x-ray projection operator, g is the measured projection signals, ∇f computes the gradient of the image, and λ is a vector of regularization coefficient, which is spatially varying for better modeling. The choice of λ is crucial for final image quality; but tuning such parameters is nontrivial. Shen et al. (2018) propose to use a DRL agent that learns a parameter-tuning policy network (PTPN) for such a tuning task. It is demonstrated that, with the aid of the agent, the final image quality reaches a level similar to that with human expert tuning.

Shen et al. (2020) propose to use DRL to learn a personalized CT scan so that the final reconstructed image quality is maximized, given a fixed dose budget. The key idea is to learn a sequential strategy that selects the acquisition angle and the needed dose for this chosen angle. The reward function is computed as

$$R(s_t, s_{t-}, a_t) = PSNR(I_t, I) - PSNR(I_{t-1}, I), \tag{32}$$

where I is the ground-truth image, I_t is the reconstructed image at time step t, and $PSNR(I', I)$ represents the Peak Signal to Noise Ratio (PSNR) value of the reconstructed image I'. Experiments are conducted using the datasets from 2016 NIH-AAPM-Mayo Clinic Low Dose CT Grand Challenge, demonstrating that the learned scanning policy yields better overall reconstruction results with the acquisition angles and dose are adaptively adjusted.

Pineda et al. (2020) propose to optimize the sequence of k-space measurements, aiming to reduce the number of measurements taken and thus accelerate the acquisition. By formulating it as a partially observable Markov decision process, a policy that maps history of k-space measurements to an index of k-space measurement to acquire next is then learned using DDQN. Similar to (32), the reward is defined as the decrease in reconstruction metric with respect to the previous reconstruction. Experiments on the fastMRI dataset of knees Zbontar et al. (2018) demonstrate that the learned policy outperforms other competing policies in terms of final reconstruction quality, over a large range of acceleration factors. Recently, Li et al. (2020) extend pixelRL Furuta et al. (2020) by assigning each pixel of the input image an agent that changes the pixel value. In their work, both reinforcement learning techniques and classical image filters are taken into to reconstruct MRI.

4.2.4. Radiotherapy planning

Radiotherapy planning often involves optimizing an objective function with constraints, which consists of multiple terms that are weighted. Weigh adjusting requires expertise from a human expert in order to yield a high quality plan. Shen et al. (2018) leverage DRL to learn a weight-tuning policy network (WTPN) that takes the current dose volume histogram of a plan as input and outputs an action that adjusts weights, with a reward function that promotes the sparing of organs at risk. The agent is then applied for planning the high-dose-rate brachytherapy for five patients, yielding the quality score 10.7% higher than human planners.

4.2.5. Video summarization

Recently, Liu et al. (2020b) introduce a fully automatic video summarization method using DRL. Their network contains an encoder-decoder CNN to first extract visual representation and then feed the feature into a Bi-LSTM to model time dependency. Finally, the RL network interprets the summarization task as a decision making process and takes actions on whether a frame should be selected for the summary set or not. In their framework, the reward is defined as the quality of the selected frames in terms of their representation, diversity, as well as the likelihood of being a standard diagnostic view plane. The proposed network can be implemented as either supervised or un-supervised manner and it obtains state-of-the-art summarization performance with highest F_1 score. The experiments are conducted based on ultrasound videos.

4.3. Miscellaneous topics

The below topics are not about analyzing clinical medical images, but they are related in general. What is common among them is that they all use reinforcement learning as a base technology.

4.3.1. Surgical gesture segmentation and classification

In a different kind of application related to medical surgery, DRL is applied to recognize surgical gestures from a video Liu and Jiang (2018). This is different from prior work that is based on graphical models Koller and Friedman (2009) such as HMM and CRF or deep learning models such as recurrent neural network and temporal convolutional network (TCN) Lea et al. (2016). In Liu and Jiang (2018), a sequential decision-making problem is set up and solved using DRL that is built upon the TCN features.

An interesting design is to use different time steps when walking through the video sequence until reaching the end. A small time step k_s is useful when the classification is not discriminative enough such as at the gesture boundaries and a large time step k_l is useful otherwise. Experimental results on the benchmark JIGSAWS dataset demonstrate that the proposed DRL achieves better performance than TCN and other competing approaches in terms of edit score due to the use of a large time step.

4.3.2. Personalized mobile health intervention

The prevalence of smartphones and wearable devices makes mobile health technology an important research direction that holds promise in impacting people's health. One idea is to use smart devices to collect and analyze raw data and to provide the device users in-time interventions, such as reduced alcohol abuse and obesity management.

Since reinforcement learning offers a sequential decision making framework, it is a natural choice for mobile data analysis. However, such an analysis often assumes that all users share the same RL model or each user has own RL model. Zhu et al. (2018) propose group-driven RL that deals with a more realistic situation: a user may be similar to some, but not all. The core idea is to find the so-called similarity network for users and cluster the users into different groups, with each group learning an RL model.

4.3.3. Computational model personalization

Computational multi-physics and multi-scale modeling Krishnamurthy et al. (2013) can improve patient stratification and therapy planning. However, the personalization of such a model, that is, the process of fitting a multi-physics computational model to clinical measurements or patient data, is a challenging research problem due to the high complexity of the models and the often noisy and sparse clinical data.

Neumann et al. (2015, 2016) propose to use an artificial agent for model personalization. Specifically, the agent learns a decision process model through exploration of the computational model offline, how the model behaves under change of parameters, and an optimal strategy for online personalization. In experiments of applying the agent to the inverse problems of cardiac electrophysiology and the personalization of a whole-body circulation model, the proposed algorithm is able to obtain equivalent results to standard methods, while being more robust (up to 11% higher success rates) and faster (up to seven times).

Finally, Abdi et al. (2018) propose to use reinforcement learning for muscle excitation estimation in biomechanical simulation. Joos et al. (2020) conduct reinforcement learning for musculoskeletal control from functional simulations.

5. Future Perspectives

DRL is a powerful framework for medical image analysis tasks. It has been successfully applied to various tasks, including image-based parameter inference in landmark localization, object detection, and registration. DRL has also been demonstrated to be an effective alternative for solving difficult optimization problems, including tuning parameters, selecting augmentation strategies, and neural architecture search. However, realizing the full potential of DRL for medical imaging requires solving several challenges ahead of us and relying on the adoptions of latest DRL advances.

5.1. Challenges ahead

Defining a reward function. It is usually hard to define or curate a learnable reward function for the task at hand because it requires the knowledge from different domains that may not always be available. A reward function with too long delay makes training difficult. In contrast, assigning a reward for each action requires careful and manual human design. Furthermore, the intermediate rewards at each time step are not always accessible. For example, an RL agent trained for a surgical suturing task would receive a feedback only after hundreds of intermediate actions. Thus, there is no feedback on how to improve the performance during the episode and what action sequences lead to the maximum final reward.

Q-learning when high-dimensional. Training a Q-function on a high-dimensional and continuous action space is challenging. For this reason, existing works using low-dimensional parameterization, typically less than 10 with an exception Krebs et al. (2017) that uses 15-D and 25-D to model 2D and 3D registration, respectively.

Data availability. DRL requires a large amount of training data or expert demonstrations. Big datasets are expensive and hard to come by. This is especially true in medical domains, partly due to strict privacy regulations and the rare nature of certain diseases. For example, retinoblastoma, the most common intraocular tumor in children, occurs with an estimated frequency of 1 in 15,000 children Ramasubramanian and Shields (2012). It is therefore challenging to collect enough data for training retinoblastoma detection and classification algorithms. Developing more data-efficient DRL algorithms is desirable to make this technology more widely applicable to the medical imaging community. Shifting from supervised to semi-supervised and unsupervised training, as well as from model-free to model-based approaches is promising directions to address the above-mentioned challenges.

Dynamic environment. Currently the approaches we have reviewed assume a stationary environment, from which observations are made. For example, the environment in the landmark detection is the image itself and what is observed is the image patch that is specified by the state (a.k.a. the location) and cropped from the image. In such case, the environment is known, but an analytic solution is not available, and DRL is used to find such an approximate solution efficiently. However, the reinforcement learning framework naturally accommodates a dynamic environment, that is, the environment itself evolves with the state and action. In other words, the only way to collect information about the environment is to interact with it. One such example is learning to scan or active acquisition Zaech et al. (2019); Zhang et al. (2019); Shen et al. (2020); Pineda et al. (2020), which opens the possibility of personalized scan with an even faster speed and at a more reduced dose. However, currently, the existing works demonstrate the idea using a simulated environment. Future works using real data from real scanning scenarios are needed.

User interaction. Another aspect worth more attention is user interaction. In the context of parametric medical image analysis, the user input essentially is an external force to escape from the local minimum trap, which gives rise to the current result. However, the subsequent behavior after escaping is largely unexplored.

Reproducibility. Reproducibility is another issue. According to Henderson et al. (2017), reproducing existing DRL work is not a straightforward task because there are non-deterministic factors even in standard benchmark environments and intrinsic variations with respect to specific methods. This statement also holds for DRL in medical imaging.

5.2. The latest DRL advances

The following latest DRL advances are worth attention and may promote new insights for many medical image analysis tasks.

Inverse DRL. DRL has been successfully applied into domains where the reward function is clearly defined. Defining such a reward function for real-world applications is challenging as it requires the knowledge from different domains that may not always be available. An example is autonomous driving, the reward function should be based on all factors such as driver's behavior, gas consumption, time, speed, safety, driving quality etc. In real-world scenario, it is hard to have control of all these factors. Different from DRL, inverse DRL Ng and Russell (2000) Abbeel and Ng (2004), a specific form of imitation learning Osa et al. (2018), infers the reward function of an agent, given its policy or observed behavior, thereby avoiding a manual specification of its reward function. In the same problem of autonomous driving, inverse RL first uses a dataset collected from the human-generated driving and then approximates the reward function for the task. Inverse RL has been successfully applied to many domains Abbeel and Ng (2004). Recently, to analyze complex human movement and control high-dimensional robot systems, Li et al. (2018) propose an online inverse RL algorithm. In You et al. (2019), both RL and inverse RL are combined to address planning problems in autonomous driving.

Multi-Agent DRL. Most of the successful DRL applications such as game Brown and Sandholm (2019), Vinyals et al. (2019), robotics Kober et al. (2013), autonomous driving Shalev-Shwartz et al. (2016), stock trading Lee et al. (2007), and social science Leibo et al. (2017) involve multiple players and require a model with multiple agents. Take autonomous driving as an instance, multi-agent DRL addresses the sequential decision-making problem which involves many autonomous agents, each of which aims to optimize its own utility return by interacting with the environment and other agents Busoniu et al. (2008). Learning in a multi-agent scenario is more difficult than a single-agent scenario because of non-stationarity Hernandez-Leal et al. (2017), multi-dimensionality Busoniu et al. (2008), credit assignment Wolpert and Tumer (2002) etc. Depending on whether the multi-agent DRL approach is either fully cooperative or fully competitive, agents can either collaborate to optimize a long-term utility or compete so that the utility is summed to zero. Recent work on Multi-Agent RL pays attention to learning a new criteria or new setup Subramanian and Mahajan (2019). There are attempts that utilize multi-agent DRL Leroy et al. (2020) Vlontzos et al. (2019) for detecting anatomical landmarks, which demonstrate that it is a relatively easy task if the agents interacts with the same environment. However, multi-agent DRL could become more challenging if the agents interact with very different environments.

Meta RL. As aforementioned, DRL algorithms consume large amounts of experience in order to learn an individual task and are unable to generalize the learned policy to newer problems. To alleviate the data challenge, Meta-RL algorithms Schweighofer and Doya (2003), Wang et al. (2016) are studied to enable agents to learn new skills from small amounts of experience. Recently there is a research interest in meta RL Nagabandi et al. (2018), Gupta et al. (2018), Sæmundsson et al. (2018), Rakelly et al. (2019), Liu et al. (2019), each using a different approach. For benchmarking and evaluation of meta RL algorithms, Yu et al. (2020) present Meta-world, which is an open-source simulator consisting of 50 distinct robotic manipulation tasks.

Imitation Learning. Imitation learning is close to learning from demonstrations which aims at training a policy to mimic an expert's behavior given the samples collected from that expert. Imitation learning is also considered as an alternative to RL/DRL to solve sequential decision-making problems. Besides inverse DRL, an imitation learning approach aforementioned, behavior cloning is another imitation learning approach to train policy under supervised learning. Stadie et al. (2017) presents a method for unsupervised third-person imitation learning to observe how other humans perform tasks. Building on top of Deep Deterministic Policy Gradients and Hindsight Experience Replay, Nair et al. (2018) propose a behavior cloning loss function to increase the level of imitating the demonstrations. Besides Q-learning, Generative Adversarial Imitation Learning Tsurumine et al. (2019) propose P-GAIL that integrate imitation learning into the policy gradient framework. P-GAIL considers both smoothness of policy update and the diversity of the learned policy by utilizing Deep P-Network Tsurumine et al. (2019).

6. Conclusions

In this paper, we present a survey of literature on the use of deep reinforcement learning in medical imaging, which demonstrates the great potential of DRL in medicine and healthcare. RL framework offers several compelling advantages compared to the traditional supervised learning approach, including i) more computationally efficient inference, ii) a smaller memory footprint or better scaling up to large image resolutions, and iii) optimal balancing between time efficiency and accuracy. The existing DRL applications for medical imaging are roughly divided into parametric medical image analysis tasks, solving optimization tasks in medical imaging, and miscellaneous applications. The remaining challenges that need to be addressed and the latest DRL advances that might promote new insights are finally discussed.

Acknowledgment

The work of Hien Van Nguyen is partly supported by the National Science Foundation (1910973).

Credit authorship contribution statement

S. Kevin Zhou: Conceptualization, Writing original draft, Writing - review and editing. **Hoang Ngan Le**: Conceptualization, Writing original draft, Writing - review and editing. **Khoa Luu**: Conceptualization, Writing original draft, Writing - review and editing. **Hien V. Nguyen**: Conceptualization, Writing original draft, Writing - review and editing. **Nicholas Ayache**: Writing - review and editing.

References

Abbeel, P., Coates, A., Ng, A.Y., 2010. Autonomous helicopter aerobatics through apprenticeship learning. The International Journal of Robotics Research 29, 1608–1639.

Abbeel, P., Ng, A.Y., 2004. Apprenticeship learning via inverse reinforcement learning, in: Proceedings of the International Conference on Machine Learning, Association for Computing Machinery. pp. 1–8.

Abdi, A.H., Saha, P., Srungarapu, P., Fels, S., 2018. Muscle excitation estimation in biomechanical simulation using naf reinforcement learning. arXiv:1809.06121 .

Akrout, M., Farahmand, A.m., Jarmain, T., Abid, L., 2019. Improving skin condition classification with a visual symptom checker trained using reinforcement learning, in: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer. pp. 549–557.

Al, W.A., Yun, I.D., 2019. Partial policy-based reinforcement learning for anatomical landmark localization in 3D medical images. IEEE Transactions on Medical Imaging 39, 1245–1255.

Alaniz, S., 2018. Deep reinforcement learning with model learning and Monte Carlo tree search in Minecraft, in: Conference on Reinforcement Learning and Decision Making.

Alansary, A., Folgoc, L.L., Vaillant, G., Oktay, O., Li, Y., Bai, W., Passerat-Palmbach, J., Guerrero, R., Kamnitsas, K., Hou, B., et al., 2018. Automatic view planning with multi-scale deep reinforcement learning agents. arXiv:1806.03228 .

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. Medical Image Analysis 53, 156–164.

Andersson, O., Heintz, F., Doherty, P., 2015. Model-based reinforcement learning in continuous environments using real-time constrained optimization, in: Proceedings of the AAAI Conference on Artificial Intelligence.

Arulkumaran, K., Deisenroth, M.P., Brundage, M., Bharath, A.A., 2017. Deep reinforcement learning: A brief survey. IEEE Signal Processing Magazine 34, 26–38.

Babaeizadeh, M., Frosio, I., Tyree, S., Clemons, J., Kautz, J., 2016. GA3C: GPU-based A3C for deep reinforcement learning. arXiv:1611.06256 .

Bae, W., Lee, S., Lee, Y., Park, B., Chung, M., Jung, K.H., 2019. Resource optimized neural architecture search for 3D medical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer. pp. 228–236.

Bagnell, J.A., Schneider, J.G., 2001. Autonomous helicopter control using reinforcement learning policy search methods, in: IEEE International Conference on Robotics and Automation (ICRA), pp. 1615–1620.

Banik, S., Rangayyan, R.M., Boag, G.S., 2009. Landmarking and segmentation of 3D CT images. Synthesis lectures on biomedical engineering 4, 1–170.

Beck, T., Bernhardt, D., Biermann, C., Dillmann, R., 2010. Validation and detection of vessel landmarks by using anatomical knowledge, in: Medical Imaging: Image Processing, International Society for Optics and Photonics. p. 76234I.

Bernstein, A., Burnaev, E., 2018. Reinforcement learning in computer vision, in: Tenth International Conference on Machine Vision, International Society for Optics and Photonics. p. 106961S.

Boedecker, J., Springenberg, J.T., Wülfing, J., Riedmiller, M., 2014. Approximate real-time optimal control based on sparse Gaussian process models, in: IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL), pp. 1–8.

Brown, N., Sandholm, T., 2019. Superhuman AI for multiplayer poker. Science 365, 885–890.

Busoniu, L., Babuska, R., De Schutter, B., 2008. A comprehensive survey of multiagent reinforcement learning. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 38, 156–172.

Cheng, K., Iriondo, C., Calivá, F., Krogue, J., Majumdar, S., Pedoia, V., 2019. Adversarial policy gradient for deep learning image augmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer. pp. 450–458.

Clarke, A.M., Friedrich, J., Tartaglia, E.M., Marchesotti, S., Senn, W., Herzog, M.H., 2015. Human and machine learning in non-markovian decision making. PLoS One 10, e0123105.

Clavera, I., Rothfuss, J., Schulman, J., Fujita, Y., Asfour, T., Abbeel, P., 2018. Model-based reinforcement learning via meta-policy optimization. arXiv:1809.05214 .

Coates, A., Abbeel, P., Ng, A.Y., 2009. Apprenticeship learning for helicopter control. Commun. ACM 52, 97–105.

Coulom, R., 2006. Efficient selectivity and backup operators in monte-carlo tree search, in: Proceedings of the 5th International Conference on Computers and Games, p. 72–83.

Crum, W.R., Hartkens, T., Hill, D., 2004. Non-rigid image registration: theory and practice. The British Journal of Radiology 77, S140–S153.

De Asis, K., Hernandez-Garcia, J.F., Holland, G.Z., Sutton, R.S., 2018. Multi-step reinforcement learning: A unifying algorithm, in: Proceedings of the AAAI Conference on Artificial Intelligence.

Deisenroth, M.P., Englert, P., Peters, J., Fox, D., 2014. Multi-task policy search for robotics, in: 2014 IEEE International Conference on Robotics and Automation (ICRA), pp. 3876–3881.

Depraetere, B., Liu, M., Pinte, G., Grondman, I., Babuska, R., 2014.

Comparison of model-free and model-based methods for time optimal hit control of a badminton robot. Mechatronics 24, 1021 – 1030.

Dou, H., Yang, X., Qian, J., Xue, W., Qin, H., Wang, X., Yu, L., Wang, S., Xiong, Y., Heng, P.A., et al., 2019. Agent with warm start and active termination for plane localization in 3D ultrasound, in: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer. pp. 290–298.

El-Fakdi, A., Carreras, M., 2008. Policy gradient based reinforcement learning for real autonomous underwater cable tracking, in: 2008 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 3635–3640.

Finn, C., Levine, S., Abbeel, P., 2016a. Guided cost learning: Deep inverse optimal control via policy optimization, in: Proceedings of the International Conference on Machine Learning, pp. 49–58.

Finn, C., Tan, X.Y., Duan, Y., Darrell, T., Levine, S., Abbeel, P., 2016b. Deep spatial autoencoders for visuomotor learning, in: Kragic, D., Bicchi, A., Luca, A.D. (Eds.), IEEE International Conference on Robotics and Automation, ICRA, pp. 512–519.

François-Lavet, V., Henderson, P., Islam, R., Bellemare, M.G., Pineau, J., 2018. An introduction to deep reinforcement learning. arXiv:1811.12560 .

Furuta, R., Inoue, N., Yamasaki, T., 2020. Pixelrl: Fully convolutional network with reinforcement learning for image processing. IEEE Transactions on Multimedia (TMM) 22, 1704–1719.

Gaon, M., Brafman, R., 2020. Reinforcement learning with non-markovian rewards, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 3980–3987.

Geist, M., Pietquin, O., 2010. A brief survey of parametric value function approximation. Rapport interne, Supélec .

Ghesu, F.C., Georgescu, B., Grbic, S., Maier, A., Hornegger, J., Comaniciu, D., 2018. Towards intelligent robust detection of anatomical structures in incomplete volumetric data. Medical Image Analysis 48, 203–213.

Ghesu, F.C., Georgescu, B., Grbic, S., Maier, A.K., Hornegger, J., Comaniciu, D., 2017. Robust multi-scale anatomical landmark detection in incomplete 3D-CT data, in: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer. pp. 194–202.

Ghesu, F.C., Georgescu, B., Mansi, T., Neumann, D., Hornegger, J., Comaniciu, D., 2016. An artificial agent for anatomical landmark detection in medical images, in: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer. pp. 229–237.

Graves, A., Mohamed, A., Hinton, G.E., 2013. Speech recognition with deep recurrent neural networks. arXiv:1303.5778 .

Gupta, A., Mendonca, R., Liu, Y., Abbeel, P., Levine, S., 2018. Meta-reinforcement learning of structured exploration strategies, in: Advances in Neural Information Processing Systems, pp. 5302–5311.

Hasselt, H.V., 2010. Double q-learning, in: Advances in Neural Information Processing Systems, pp. 2613–2621.

Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., Meger, D., 2017. Deep reinforcement learning that matters. arXiv:1709.06560 .

Hernandez-Leal, P., Kaisers, M., Baarslag, T., de Cote, E.M., 2017. A survey of learning in multiagent environments: Dealing with non-stationarity. arXiv:1707.09183 .

Hester, T., Quinlan, M., Stone, P., 2011. A real-time model-based reinforcement learning architecture for robot control. arXiv:1105.1749 .

Holliday, J.B., Le, N.T., 2020. Follow then forage exploration: Improving asynchronous advantage actor critic. International Conference on Soft Computing, Artificial Intelligence and Applications (SAI) , 107–118.

Huang, Y., Yang, X., Li, R., Qian, J., Huang, X., Shi, W., Dou, H., Chen, C., Zhang, Y., Luo, H., Frangi, A., Xiong, Y., Ni, D., 2020a. Searching collaborative agents for multi-plane localization in 3D ultrasound, in: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI).

Huang, Y., Yang, X., Li, R., Qian, J., Huang, X., Shi, W., Dou, H., Chen, C., Zhang, Y., Luo, H., et al., 2020b. Searching collaborative agents for multi-plane localization in 3D ultrasound, in: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer. pp. 553–562.

Isensee, F., Petersen, J., Klein, A., Zimmerer, D., Jaeger, P.F., Kohl, S., Wasserthal, J., Koehler, G., Norajitra, T., Wirkert, S., et al., 2018. nnU-Net: Self-adapting framework for u-net-based medical image segmentation. arXiv:1809.10486 .

Jia, Z., Yang, L., Szepesvari, C., Wang, M., 2020. Model-based reinforcement learning with value-targeted regression, in: Proceedings of the 2nd Conference on Learning for Dynamics and Control, The Cloud. pp. 666–686.

Jie, Z., Liang, X., Feng, J., Jin, X., Lu, W., Yan, S., 2016. Tree-structured reinforcement learning for sequential object localization, in: Advances in Neural Information Processing Systems, pp. 127–135.

Johnson, H.J., Christensen, G.E., 2002. Consistent landmark and intensity-based image registration. IEEE Transactions on Medical Imaging 21, 450–461.

Joos, E., Péan, F., Goksel, O., 2020. Reinforcement learning of musculoskeletal control from functional simulations, in: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer. pp. 135–145.

Klein, S., Staring, M., Murphy, K., Viergever, M.A., Pluim, J.P., 2010. Elastix: a toolbox for intensity-based medical image registration. IEEE Transactions on Medical Imaging 29, 196–205.

Kober, J., Bagnell, J.A., Peters, J., 2013. Reinforcement learning in robotics: A survey. The International Journal of Robotics Research 32, 1238–1274.

Kober, J., Peters, J., 2014. Policy search for motor primitives in robotics, in: Learning Motor Skills. Springer, pp. 83–117.

Koller, D., Friedman, N., 2009. Probabilistic graphical models: principles and techniques. MIT Press.

Konda, V.R., Tsitsiklis, J.N., 2000. Actor-critic algorithms, in: Advances in Neural Information Processing Systems, pp. 1008–1014.

Krebs, J., Mansi, T., Delingette, H., Zhang, L., Ghesu, F.C., Miao, S., Maier, A.K., Ayache, N., Liao, R., Kamen, A., 2017. Robust non-rigid registration through agent-based action learning, in: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer. pp. 344–352.

Krishnamurthy, A., Villongco, C.T., Chuang, J., Frank, L.R., Nigam, V., Belezzuoli, E., Stark, P., Krummen, D.E., Narayan, S., Omens, J.H., et al., 2013. Patient-specific models of cardiac biomechanics. Journal of Computational Physics 244, 4–21.

Kupcsik, A., Deisenroth, M., Peters, J., Neumann, G., 2013. Data-efficient generalization of robot skills with contextual policy search, in: Proceedings of the AAAI Conference on Artificial Intelligence.

Kupcsik, A., Deisenroth, M.P., Peters, J., Loh, A.P., Vadakkepat, P., Neumann, G., 2017. Model-based contextual policy search for data-efficient generalization of robot skills. Artificial Intelligence 247, 415–439.

Kurutach, T., Clavera, I., Duan, Y., Tamar, A., Abbeel, P., 2018. Model-ensemble trust-region policy optimization .

Lagoudakis, M.G., 2017. Value Function Approximation. Springer US, Boston, MA. pp. 1311–1323.

Lay, N., Birkbeck, N., Zhang, J., Zhou, S.K., 2013. Rapid multi-organ segmentation using context integration and discriminative models, in: International Conference on Information Processing in Medical Imaging, Springer. pp. 450–462.

Lea, C., Vidal, R., Reiter, A., Hager, G.D., 2016. Temporal convolutional networks: A unified approach to action segmentation, in: European Conference on Computer Vision, Springer. pp. 47–54.

Lee, J.W., Park, J., O, J., Lee, J., Hong, E., 2007. A multiagent approach to q-learning for daily stock trading. Trans. Sys. Man Cyber. Part A 37, 864–877.

Lee, K., Kim, S.A., Choi, J., Lee, S.W., 2018. Deep reinforcement learning in continuous action spaces: a case study in the game of simulated curling, in: Dy, J., Krause, A. (Eds.), Proceedings of the International Conference on Machine Learning, PMLR. pp. 2937–2946.

Leibo, J.Z., Zambaldi, V.F., Lanctot, M., Marecki, J., Graepel, T., 2017. Multi-agent reinforcement learning in sequential social dilemmas. arXiv:1702.03037 .

Leroy, G., Rueckert, D., Alansary, A., 2020. Communicative reinforcement learning agents for landmark detection in brain images, in: Machine Learning in Clinical Neuroimaging and Radiogenomics in Neuro-oncology. Springer, pp. 177–186.

Levine, S., Koltun, V., 2014. Learning complex neural network policies with trajectory optimization, in: Proceedings of the International Conference on Machine Learning, pp. 829–837.

Li, K., Rath, M., Burdick, J.W., 2018. Inverse reinforcement learning via function approximation for clinical motion analysis, in: 2018 IEEE International Conference on Robotics and Automation (ICRA), pp. 610–617.

Li, W., Feng, X., An, H., Ng, X.Y., Zhang, Y.J., 2020. MRI reconstruction with interpretable pixel-wise operations using reinforcement learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 792–799.

Li, Y., 2017. Deep reinforcement learning: An overview. arXiv:1701.07274 .

Liao, R., Miao, S., de Tournemire, P., Grbic, S., Kamen, A., Mansi, T., Comaniciu, D., 2017. An artificial agent for robust image registration., in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 4168–4175.

Liao, X., Li, W., Xu, Q., Wang, X., Jin, B., Zhang, X., Wang, Y., Zhang, Y., 2020. Iteratively-refined interactive 3D medical image segmentation with multi-agent reinforcement learning, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR), pp. 9394–9402.

Lillicrap, T.P., Hunt, J.J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., Wierstra, D., 2015. Continuous control with deep reinforcement learning. arXiv:1509.02971 .

Lillicrap, T.P., Hunt, J.J., Pritzel, A.e., Heess, N., Erez, T., Tassa, Y., Silver, D., Wierstra, D., 2015. Continuous control with deep reinforcement learning. arXiv e-prints , arXiv:1509.02971arXiv:1509.02971.

Liu, D., Jiang, T., 2018. Deep reinforcement learning for surgical gesture segmentation and classification. arXiv:1806.08089 .

Liu, D., Zhou, K.S., Bernhardt, D., Comaniciu, D., 2010. Search strategies for multiple landmark detection by submodular maximization, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR), IEEE. pp. 2831–2838.

Liu, H., Simonyan, K., Yang, Y., 2018. Darts: Differentiable architecture search. arXiv:1806.09055 .

Liu, H., Socher, R., Xiong, C., 2019. Taming maml: Efficient unbiased meta-reinforcement learning, in: Proceedings of the International Conference on Machine Learning, pp. 4061–4071.

Liu, T., Meng, Q., Vlontzos, A., Tan, J., Rueckert, D., Kainz, B., 2020a. Ultrasound video summarization using deep reinforcement learning, in: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer. pp. 483–492.

Liu, T., Meng, Q., Vlontzos, A., Tan, J., Rueckert, D., Kainz, B., 2020b. Ultrasound video summarization using deep reinforcement learning, in: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 483–492.

Liu, W., Liu, F., Tang, R., Liao, B., Chen, G., Heng, P.A., 2020c. Balancing between accuracy and fairness for interactive recommendation with reinforcement learning, in: Lauw, H.W., Wong, R.C.W., Ntoulas, A., Lim, E.P., Ng, S.K., Pan, S.J. (Eds.), Advances in Knowledge Discovery and Data Mining, Springer International Publishing, Cham. pp. 155–167.

Luketina, J., Nardelli, N., Farquhar, G., Foerster, J., Andreas, J., Grefenstette, E., Whiteson, S., Rocktäschel, T., 2019. A survey of reinforcement learning informed by natural language. arXiv:1906.03926 .

Luo, G., Dong, S., Wang, K., Zhang, D., Gao, Y., Chen, X., Zhang, H., Li, S., 2019. A deep reinforcement learning framework for frame-by-frame plaque tracking on intravascular optical coherence tomography image, in: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer. pp. 12–20.

Ma, J., Zhu, X., Yang, D., Chen, J., Wu, G., 2020. Attention-guided deep graph neural network for longitudinal Alzheimer's disease analysis, in: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer. pp. 387–396.

Ma, K., Wang, J., Singh, V., Tamersoy, B., Chang, Y.J., Wimmer, A., Chen, T., 2017. Multimodal image registration with deep context reinforcement learning, in: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer. pp. 240–248.

Maicas, G., Bradley, A.P., Nascimento, J.C., Reid, I., Carneiro, G., 2019. Pre and post-hoc diagnosis and interpretation of malignancy from breast DCE-MRI. Medical Image Analysis 58, 101562.

Maicas, G., Carneiro, G., Bradley, A.P., Nascimento, J.C., Reid, I., 2017. Deep reinforcement learning for active breast lesion detection from DCE-MRI, in: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer. pp. 665–673.

Majeed, S.J., Hutter, M., 2018. On q-learning convergence for non-markov decision processes., in: IJCAI, pp. 2546–2552.

Martinez-Marin, T., Duckett, T., 2005. Fast reinforcement learning for vision-guided mobile robots, in: Proceedings of the 2005 IEEE International Conference on Robotics and Automation, pp. 4170–4175.

Masson, W., Ranchod, P., Konidaris, G., 2016. Reinforcement learning with parameterized actions, in: Proceedings of the AAAI Conference on Artificial Intelligence, AAAI Press. pp. 1934–1940.

Mnih, V., Badia, A.P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., Kavukcuoglu, K., 2016. Asynchronous methods for deep reinforcement learning, in: Proceedings of the International Conference on Machine Learning, 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:1312.5602 .

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

Mordatch, I., Mishra, N., Eppner, C., Abbeel, P., 2016. Combining model-based policy search with online model learning for control of physical humanoids, in: 2016 IEEE International Conference on Robotics and Automation (ICRA), pp. 242–248.

Morimoto, J., Atkeson, C.G., 2009. Nonparametric representation of an approximated poincaré map for learning biped locomotion, in: Autonomous Robots, p. 131–144.

Morimoto, J., Zeglin, G., Atkeson, C.G., 2003. Minimax differential dynamic programming: application to a biped walking robot, in: Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003), pp. 1927–1932.

Nachum, O., Norouzi, M., Xu, K., Schuurmans, D., 2017. Bridging the gap between value and policy based reinforcement learning, in: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc.

Nagabandi, A., Clavera, I., Liu, S., Fearing, R.S., Abbeel, P., Levine, S., Finn, C., 2018. Learning to adapt in dynamic, real-world environments through meta-reinforcement learning. arXiv:1803.11347 .

Nair, A., McGrew, B., Andrychowicz, M., Zaremba, W., Abbeel, P., 2018. Overcoming exploration in reinforcement learning with demonstrations, in: 2018 IEEE International Conference on Robotics and Automation (ICRA), pp. 6292–6299. doi:10.1109/ICRA.2018.8463162.

Neumann, D., Mansi, T., Itu, L., Georgescu, B., Kayvanpour, E., Sedaghat-Hamedani, F., Amr, A., Haas, J., Katus, H., Meder, B., et al., 2016. A self-taught artificial agent for multi-physics computational model personalization. Medical Image Analysis 34, 52–64.

Neumann, D., Mansi, T., Itu, L., Georgescu, B., Kayvanpour, E., Sedaghat-Hamedani, F., Haas, J., Katus, H., Meder, B., Steidl, S., et al., 2015. Vito–a generic agent for multi-physics model personalization: Application to heart modeling, in: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer. pp. 442–449.

Ng, A.Y., Russell, S.J., 2000. Algorithms for inverse reinforcement learning, in: Proceedings of the International Conference on Machine Learning, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. p. 663–670.

Nguyen, T.T., Li, Z., Silander, T., Leong, T.Y., 2013. Online feature selection for model-based reinforcement learning, in: Proceedings of the International Conference on Machine Learning, p. I–498–I–506.

Osa, T., Pajarinen, J., Neumann, G., Bagnell, J.A., Abbeel, P., Peters, J., 2018. An Algorithmic Perspective on Imitation Learning. doi:10.1561/2300000053.

Pauly, O., Glocker, B., Criminisi, A., Mateus, D., Möller, A.M., Nekolla, S., Navab, N., 2011. Fast multiple organ detection and localization in whole-body MR Dixon sequences, in: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer. pp. 239–247.

Pesce, E., Withey, S.J., Ypsilantis, P.P., Bakewell, R., Goh, V., Montana, G., 2019. Learning to detect chest radiographs containing pulmonary lesions using visual attention networks. Medical Image Analysis 53, 26–38.

Peters, J., Bagnell, J.A., 2010. Policy Gradient Methods. Springer US, Boston, MA. pp. 774–776.

Peters, J., Schaal, S., 2008. Natural actor-critic. Neurocomputing 71, 1180–1190.

Pineda, L., Basu, S., Romero, A., Calandra, R., Drozdzal, M., 2020. Active mr k-space sampling with reinforcement learning, in: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer. pp. 23–33.

Plaat, A., Kosters, W., Preuss, M., 2020. Deep model-based reinforcement learning for high-dimensional problems, a survey. arXiv:2008.05598.

Qaiser, T., Rajpoot, N.M., 2019. Learning where to see: A novel attention model for automated immunohistochemical scoring. IEEE Transactions on Medical Imaging 38, 2620–2631.

Qin, T., Wang, Z., He, K., Shi, Y., Gao, Y., Shen, D., 2020. Automatic data augmentation via deep reinforcement learning for effective kidney tumor segmentation, in: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 1419–1423.

Qin, T., Wang, Z., He, K., Shi, Y., Gao, Y., Shen, D., 2020. Automatic data augmentation via deep reinforcement learning for effective kidney tumor segmentation, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1419–1423.

Rakelly, K., Zhou, A., Finn, C., Levine, S., Quillen, D., 2019. Efficient off-policy meta-reinforcement learning via probabilistic context variables, in: Proceedings of the International Conference on Machine Learning, pp. 5331–5340.

Ramasubramanian, A., Shields, C.L., 2012. Retinoblastoma. JP Medical Ltd.

Ravishankar, H., Venkataramani, R., Thiruvenkadam, S., Sudhakar, P., Vaidya, V., 2017. Learning and incorporating shape models for semantic segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 203–211.

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 (MICCAI), Springer. pp. 234–241.

Rudin, L.I., Osher, S., Fatemi, E., 1992. Nonlinear total variation based noise removal algorithms. Physica D: nonlinear phenomena 60, 259–268.

Rummery, G.A., Niranjan, M., 1994. On-line Q-learning using connectionist systems. volume 37. University of Cambridge, Department of Engineering Cambridge, UK.

Sæmundsson, S., Hofmann, K., Deisenroth, M.P., 2018. Meta reinforcement learning with latent variable Gaussian processes. arXiv:1803.07551 .

Sahba, F., Tizhoosh, H.R., Salama, M.M., 2006. A reinforcement learning framework for medical image segmentation, in: IJCNN, pp. 511–517.

Schulman, J., Levine, S., Moritz, P., Jordan, M.I., Abbeel, P., 2015. Trust Region Policy Optimization. arXiv:1502.05477 .

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O., 2017. Proximal Policy Optimization Algorithms. arXiv:1707.06347, keywords = Computer Science - Machine Learning, .

Schweighofer, N., Doya, K., 2003. Meta-learning in reinforcement learning. Neural Networks 16, 5–9.

Seifert, S., Kelm, M., Moeller, M., Mukherjee, S., Cavallaro, A., Huber, M., Comaniciu, D., 2010. Semantic annotation of medical images, in: Medical Imaging 2010: Advanced PACS-based Imaging Informatics and Therapeutic Applications, International Society for Optics and Photonics. p. 762808.

Shalev-Shwartz, S., Shammah, S., Shashua, A., 2016. Safe, multi-agent, reinforcement learning for autonomous driving. arXiv:1610.03295 .

Shen, C., Gonzalez, Y., Chen, L., Jiang, S.B., Jia, X., 2018. Intelligent parameter tuning in optimization-based iterative ct reconstruction via deep reinforcement learning. IEEE Transactions on Medical Imaging 37, 1430–1439.

Shen, C., Gonzalez, Y., Klages, P., Qin, N., Jung, H., Chen, L., Nguyen, D., Jiang, S.B., Jia, X., 2019. Intelligent inverse treatment planning via deep reinforcement learning, a proof-of-principle study in high dose-rate brachytherapy for cervical cancer. Physics in Medicine & Biology 64, 115013.

Shen, Z., Wang, Y., Wu, D., Yang, X., Dong, B., 2020. Learning to scan: A deep reinforcement learning approach for personalized scanning in ct imaging. arXiv:2006.02420 .

Shokri, M., Tizhoosh, H.R., 2003. Using reinforcement learning for image thresholding, in: Electrical and Computer Engineering, 2003. IEEE CCECE 2003. Canadian Conference on, IEEE. pp. 1231–1234.

Stadie, B.C., Abbeel, P., Sutskever, I., 2017. Third-person imitation learning. arXiv:1703.01703 .

Subramanian, J., Mahajan, A., 2019. Reinforcement learning in stationary mean-field games, International Foundation for Autonomous Agents and Multiagent Systems. p. 251–259.

Sutton, R.S., Barto, A.G., 2018. Reinforcement learning: An introduction. MIT Press.

Sutton, R.S., McAllester, D.A., Singh, S.P., Mansour, Y., 2000. Policy gradient methods for reinforcement learning with function approximation, in: Advances in Neural Information Processing Systems 12, pp. 1057–1063.

Torr, P.H., Zisserman, A., 2000. Mlesac: A new robust estimator with application to estimating image geometry. Computer Vision and Image Understanding 78, 138–156.

Tsurumine, Y., Cui, Y., Uchibe, E., Matsubara, T., 2019. Deep reinforcement learning with smooth policy update: Application to robotic cloth manipulation. Robotics and Autonomous Systems 112, 72 – 83.

Tsurumine, Y., Cui, Y., Yamazaki, K., Matsubara, T., 2019. Generative adversarial imitation learning with deep p-network for robotic cloth manipulation, in: 2019 IEEE-RAS 19th International Conference on Humanoid Robots (Humanoids), pp. 274–280. doi:10.1109/Humanoids43949.2019.9034991.

van Hasselt, H., Guez, A., Silver, D., 2015. Deep reinforcement learning with double q-learning. arXiv:1509.06461 .

Vinyals, O., Babuschkin, I., Chung, J., Mathieu, M., Jaderberg, M., Czarnecki, W., Dudzik, A., Huang, A., Georgiev, P., Powell, R., Ewalds, T., Horgan, D., Kroiss, M., Danihelka, I., Agapiou, J., Oh, J., Dalibard, V., Choi, D., Sifre, L., Sulsky, Y., Vezhnevets, S., Molloy, J., Cai, T., Budden, D., Paine, T., Gulcehre, C., Wang, Z., Pfaff, T., Pohlen, T., Yogatama, D., Cohen, J., McKinney, K., Smith, O., Schaul, T., Lillicrap, T., Apps, C., Kavukcuoglu, K., Hassabis, D., Silver, D., 2019. AlphaStar: Mastering the Real-Time Strategy Game StarCraft II. https://deepmind.com/blog/alphastar-mastering-real-time-strategy-game-starcraft-ii/.

Vlontzos, A., Alansary, A., Kamnitsas, K., Rueckert, D., Kainz, B., 2019. Multiple landmark detection using multi-agent reinforcement learning, in: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer. pp. 262–270.

Wang, J., Miao, J., Yang, X., Li, R., Zhou, G., Huang, Y., Lin, Z., Xue, W., Jia, X., Zhou, J., et al., 2020. Auto-weighting for breast cancer classification in multimodal ultrasound, in: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer. pp. 190–199.

Wang, J.X., Kurth-Nelson, Z., Tirumala, D., Soyer, H., Leibo, J.Z., Munos, R., Blundell, C., Kumaran, D., Botvinick, M., 2016. Learning to reinforcement learn. arXiv:1611.05763 .

Wang, L., Lekadir, K., Lee, S.L., Merrifield, R., Yang, G.Z., 2013. A general framework for context-specific image segmentation using reinforcement learning. IEEE Transactions on Medical Imaging 32, 943–956.

Wang, T., Bao, X., Clavera, I., Hoang, J., Wen, Y., Langlois, E., Zhang, S., Zhang, G., Abbeel, P., Ba, J., 2019. Benchmarking model-based reinforcement learning. arXiv:1907.02057 .

Wang, Z., Schaul, T., Hessel, M., Van Hasselt, H., Lanctot, M., De Freitas, N., 2015. Dueling network architectures for deep reinforcement learning. arXiv:1511.06581 .

Watkins, C.J., Dayan, P., 1992. Q-learning. Machine Learning 8, 279–292.

Williams, R.J., 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine Learning 8, 229–256.

Wilson, A., Fern, A., Tadepalli, P., 2014. Using trajectory data to improve bayesian optimization for reinforcement learning. Journal of Machine Learning Research 15, 253–282.

Wolpert, D.H., Tumer, K., 2002. Optimal payoff functions for members of collectives, in: Modeling complexity in economic and social systems. World Scientific, pp. 355–369.

Xu, B., Liu, J., Hou, X., Liu, B., Garibaldi, J., Ellis, I.O., Green, A., Shen, L., Qiu, G., 2019. Attention by selection: A deep selective attention approach to breast cancer classification. IEEE Transactions on Medical Imaging 39, 1930–1941.

Xu, Z., Huang, Q., Park, J., Chen, M., Xu, D., Yang, D., Liu, D., Zhou, S.K., 2017. Supervised action classifier: Approaching landmark detection as image partitioning, in: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer. pp. 338–346.

Yang, D., Roth, H., Xu, Z., Milletari, F., Zhang, L., Xu, D., 2019. Searching learning strategy with reinforcement learning for 3D medical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer. pp. 3–11.

Yang, H., Shan, C., Kolen, A.F., de With, P.H.N., 2020a. Deep q-network-driven catheter segmentation in 3D US by hybrid constrained semi-supervised learning and dual-unet, in: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 646–655.

Yang, H., Shan, C., Kolen, A.F., et al., 2020b. Deep q-network-driven catheter segmentation in 3D US by hybrid constrained semi-supervised learning and dual-unet, in: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer. pp. 646–655.

Yang, X., Yu, L., Li, S., Wen, H., Luo, D., Bian, C., Qin, J., Ni, D., Heng, P., 2019. Towards automated semantic segmentation in prenatal volumetric ultrasound. IEEE Transactions on Medical Imaging 38, 180–193.

Ye, J., Xue, Y., Long, L.R., Antani, S., Xue, Z., Cheng, K.C., Huang, X., 2020. Synthetic sample selection via reinforcement learning, in: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer. pp. 53–63.

You, C., Lu, J., Filev, D., Tsiotras, P., 2019. Advanced planning for autonomous vehicles using reinforcement learning and deep inverse reinforcement learning. Robotics and Autonomous Systems 114, 1–18.

Yu, T., Quillen, D., He, Z., Julian, R., Hausman, K., Finn, C., Levine, S., 2020. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning, in: Conference on Robot Learning, pp. 1094–1100.

Zaech, J.N., Gao, C., Bier, B., Taylor, R., Maier, A., Navab, N., Unberath, M., 2019. Learning to avoid poor images: Towards task-aware c-arm cone-beam ct trajectories, in: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer. pp. 11–19.

Zbontar, J., Knoll, F., Sriram, A., Muckley, M.J., Bruno, M., Defazio, A., Parente, M., Geras, K.J., Katsnelson, J., Chandarana, H., et al., 2018. fastMRI: An open dataset and benchmarks for accelerated MRI. arXiv:1811.08839 .

Zha, D., Lai, K.H., Zhou, K., Hu, X., 2019. Experience replay optimization. arXiv:1906.08387 .

Zhang, M., Xu, J., Abaci Turk, E., Grant, P.E., Golland, P., Adalsteinsson, E., 2020a. Enhanced detection of fetal pose in 3D MRI by deep reinforcement learning with physical structure priors on anatomy, in: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 396–405.

Zhang, P., Wang, F., Zheng, Y., 2018. Deep reinforcement learning for vessel centerline tracing in multi-modality 3D volumes, in: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer. pp. 755–763.

Zhang, Y., Amadou, A.a., Voigt, I., Mihalef, V., Houle, H., John, M., Mansi, T., Liao, R., 2020b. A bottom-up approach for real-time mitral valve annulus modeling on 3D echo images, in: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer. pp. 458–467.

Zhang, Y., Luo, G., Wang, W., Wang, K., 2020c. Branch-aware double DQN for centerline extraction in coronary CT angiography, in: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer. pp. 35–44.

Zhang, Y., Luo, G., Wang, W., Wang, K., 2020d. Branch-aware double dqn for centerline extraction in coronary ct angiography, in: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 35–44.

Zhang, Z., Romero, A., Muckley, M.J., Vincent, P., Yang, L., Drozdzal, M., 2019. Reducing uncertainty in undersampled MRI reconstruction with active acquisition, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR), pp. 2049–2058.

Zhou, S.K., 2010. Shape regression machine and efficient segmentation of left ventricle endocardium from 2d b-mode echocardiogram. Medical Image Analysis 14, 563–581.

Zhou, S.K., Greenspan, H., Davatzikos, C., Duncan, J.S., van Ginneken, B., Madabhushi, A., Prince, J.L., Rueckert, D., Summers, R.M., 2020. A review of deep learning in medical imaging: Image traits, technology trends, case studies with progress highlights, and future promises. arXiv:2008.09104 .

Zhu, F., Guo, J., Xu, Z., Liao, P., Yang, L., Huang, J., 2018. Group-driven reinforcement learning for personalized mhealth intervention, in: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer. pp. 590–598.

Zoph, B., Le, Q.V., 2016. Neural architecture search with reinforcement learning. arXiv:1611.01578 .