Deep Generative Adversarial Reinforcement Learning for Semi-Supervised Segmentation of Low-Contrast and Small Objects in Medical Images

Chenchu Xu, Tong Zhang , Dong Zhang, Dingwen Zhang, and Junwei Han., *Fellow, IEEE*

***Abstract*— Deep reinforcement learning (DRL) has demonstrated impressive performance in medical image segmentation, particularly for low-contrast and small medical objects. However, current DRL-based segmentation methods face limitations due to the optimization of error propagation in two separate stages and the need for a significant amount of labeled data. In this paper, we propose a novel deep generative adversarial reinforcement learning (DGARL) approach that, for the first time, enables end-to-end semi-supervised medical image segmentation in the DRL domain. DGARL ingeniously establishes a pipeline that integrates DRL and generative adversarial networks (GANs) to optimize both detection and segmentation tasks holistically while mutually enhancing each other. Specifically, DGARL introduces two innovative components to facilitate this integration in semi-supervised settings. First, a task-joint GAN with two discriminators links the detection results to the GAN's segmentation performance evaluation, allowing simultaneous joint evaluation and feedback. This ensures that DRL and GAN can be directly optimized based on each other's results. Second, a bidirectional exploration DRL integrates backward exploration and forward exploration to ensure the DRL agent explores the correct direction when forward exploration is disabled due to lack of explicit rewards. This mitigates the issue of unlabeled data being unable to provide rewards and rendering DRL unexplorable. Comprehensive experiments on three generalization datasets, comprising a total of 640 patients, demonstrate that our novel DGARL achieves 85.02% Dice and improves at least 1.91% for brain tumors, achieves 73.18% Dice and improves at least 4.28% for liver tumors, and achieves 70.85% Dice and improves at least 2.73% for pancreas compared to the ten most recent advanced methods, our results attest to the superiority of DGARL. Code is available at GitHub.**

***Index Terms*— Medical Image Segmentation, Deep Reinforcement Learning(DRL), Generative Adversarial Networks(GANs).**

This paper was submitted for review on June 3, 2023. This work was supported in part by The National Natural Science Foundation of China (62106001, U1908211), The University Synergy Innovation Program of Anhui Province under Grant (GXXT-2021-007), and The Anhui Provincial Natural Science Foundation (2208085Y19). (Corresponding author: D. Zhang.)

C. Xu is with the School of Computer Science and Technology, Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, Anhui University, Hefei, China, and the Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei, China. (e-mail: cxu332@gmail.com).

T. Zhang is with the School of Computer Science and Technology, Anhui University, Hefei, China. (E21201034@stu.ahu.edu.cn).

D. Zhang is with the Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, Canada. (e-mail: zhangdong9612@gmail.com).

D. Zhang and J. Han are with the School of Automation, Northwestern Polytechnical University, Xi'an, China. (zhangdingwen2006yyy@gmail.com; junweihan2010@gmail.com).

I. Introduction

DEEP reinforcement learning (DRL) has shown promising results in the field of medical image segmentation, particularly in the segmentation of low-contrast and small medical objects such as landmarks and lesions [1], [3], [4]. Unlike direct mapping-based deep learning methods, DRL approaches the task by treating the input image as an environment and deploying a self-learning agent that iteratively explores the environment to learn to make sequential optimal decisions. DRL's iterative exploration mechanism [23] significantly increases the learnable feature space of desired objects and amplifies the feature differences in different regions on the target boundary. This process effectively avoids feature blurring caused by unclear boundaries, leading to improved performance in low-contrast images. Additionally, DRL's sequential decision mechanism allows the investigation of every state in the environment and treats the possibility of detecting each target equally. This approach ensures that small targets are not neglected due to unbalanced samples [1]. Overall, DRL shows great potential for improving medical image segmentation, especially in cases where traditional deep learning methods struggle.

Despite the advantages of DRL for medical image segmentation highlighted in the previous paragraph, recent DRL-based methods for medical image segmentation [1], [3], [4] face two well-known challenges. Firstly, as shown in Fig. 1, existing methods typically consist of two separate stages: a coarse detection stage and a fine segmentation stage, where the segmentation stage relies on the detection result as a prior. This approach can result in inaccurate segmentation due to the propagation of detection errors to the subsequent segmentation stage, which compromises the overall performance. Specifically, the one-way propagation between the two stages makes it difficult to optimize the DRL network based on segmentation

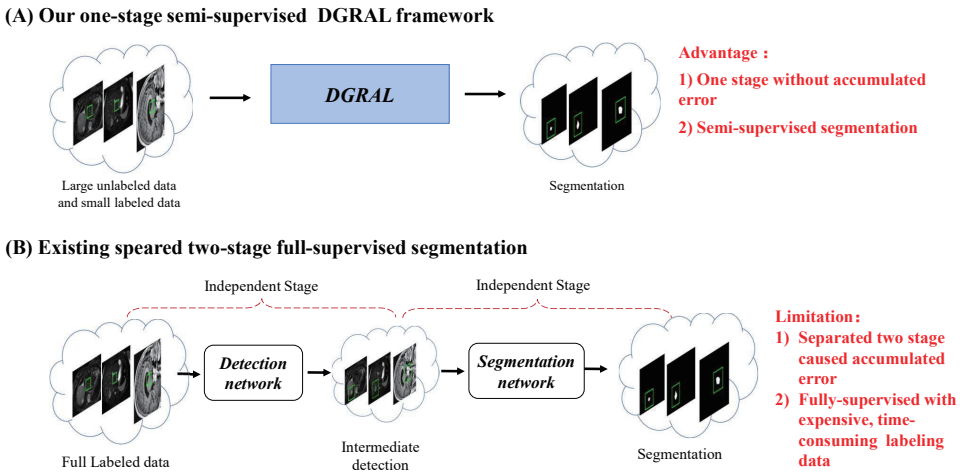

Fig. 1: Our DGARL pioneers end-to-end training in semi-supervised medical image segmentation, unlike recent DRL-based methods that rely on time-consuming, fully supervised learning and a two-stage process. This conventional approach often leads to segmentation inaccuracies due to error propagation between stages.

errors. Secondly, existing DRL-based methods for medical image segmentation rely on fully supervised learning, which requires a significant amount of fully-annotated data to achieve satisfactory performance. However, manually annotating medical images is often time-consuming and expensive. In addition, fully annotated data is usually not required for clinical practice and may not be easily available for researchers to access, making it challenging to apply these methods to new or rare categories. Overall, while DRL shows promise for medical image segmentation, these challenges need to be addressed to improve the accuracy and scalability of DRL-based methods in clinical practice and research.

To address the aforementioned challenges, this paper proposes a novel deep generative adversarial reinforcement learning (DGARL), for the first time, to enable end-to-end semi-supervised medical image segmentation by using deep reinforcement learning. Unlike existing methods, which rely on a two-stage process, our DGARL seamlessly integrates its designed DRL component and Generative Adversarial Networks (GANs) component into a single reciprocal pipeline, where DRL and GAN can promote each other. Specifically, as shown in Fig. 2, the DRL component estimates the potential region of objects and guides the GAN component to perform segmentation. Simultaneously, the GAN component evaluates the segmentation performance and provides feedback to guide the DRL component. By integrating DRL and GAN into a single pipeline, we are able to train both components simultaneously, optimize their errors jointly, and avoid accumulative errors. Additionally, our approach enables GAN to perform semi-supervised segmentation effectively by leveraging the estimated regions from DRL to segment objects from unlabeled data. Overall, the proposed DGARL framework overcomes the limitations of existing DRL-based medical image segmentation methods and allows end-to-end semi-supervised segmentation. Our approach enables DRL and GAN to promote each other, leading to better segmentation results.

However, incorporating DRL and GAN seamlessly in a semi-supervised segmentation setting is an unprecedented and incredibly challenging task, for several reasons. Firstly, DRL and GAN perform different tasks, making it difficult to optimize their cooperation. Specifically, GAN segmentation cannot directly provide rewards for DRL detection [16], and conversely, DRL detection cannot directly drive GAN segmentation. During DRL's initial exploration stage, when the detection bounding box from DRL is far from the segmented object, GAN cannot perform segmentation as there is no object to segment, and therefore cannot provide any reward. Secondly, the semi-supervised segmentation setting exacerbates the challenges of optimizing cooperation between DRL and GAN, as GAN cannot evaluate unlabeled datasets as quantitative rewards to drive DRL [17]. The principle of semi-supervised GAN is to qualitatively evaluate whether its segmentation is consistent with labeled data distribution, rather than how good or bad it is. On the other hand, DRL requires accurate quantitative rewards to compare different explored actions and find an optimal detection policy. If using such a qualitative evaluation from GAN as the rewards, DRL exploration is rendered ineffective [18].

Therefore, our DGARL further proposes two novities. First, a task-joint GAN architecture that enables joint evaluation of both segmentation and detection tasks, allowing for direct optimization of both DRL and GAN through each other's task results. Specifically, the task-joint GAN utilizes two discriminators that collaboratively perform segmentation and detection tasks, with the detection results being linked to the GAN's segmentation performance evaluation. Such a joint evaluation on both tasks provides feedback to both GAN and DRL simultaneously, enabling adaptive optimization according to the different task requirements and schedules. Secondly, a bidirectional exploration DRL combines both forward and backward exploration directions. This approach ensures that when the agent's forward exploration cannot be carried out due to a lack of explicit rewards, the backward exploration can guide the agent in the correct direction. As a result, bidirectional exploration DRL can explore under qualitative reward in semi-supervised learning and significantly improve the effectiveness of detection. These two novelties allow for effective collaboration optimization between DRL and GAN, even in challenging semi-supervised segmentation settings.

In summary, the contributions of this work include:

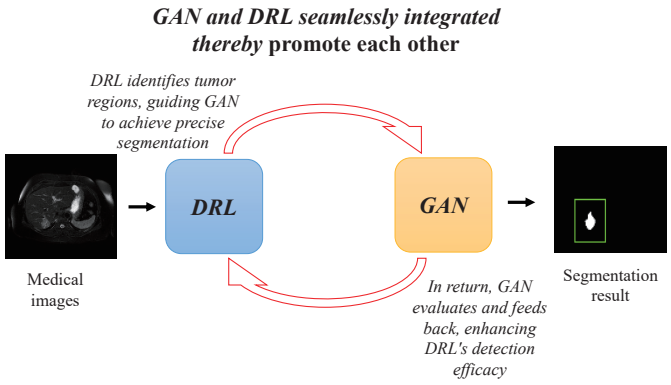

Fig. 2: In our DGARL system, DRL and GAN are integrated into a mutual enhancement pipeline, where both components have the ability to boost the other's performance. Specifically, DRL estimates the prospective object regions, thereby directing the GAN to perform segmentation tasks. Concurrently, GAN assesses the quality of segmentation and offers valuable feedback that refines the trajectory of the DRL component.

- For the first time, DRL has power in semi-supervised segmentation for medical images as an end-to-end training. By providing a solution to the challenge of obtaining reliable rewards from unlabeled data, we have overcome a major obstacle to optimizing DRL in this context.
- A pioneering framework seamlessly integrates GAN and DRL, with a reciprocal pipeline that allows for seamless collaboration between the DRL's detection task and GAN's segmentation task. This novelty leads to greatly advancing the performance of decision-making tasks for segmentation tasks.
- A task-joint adversarial training architecture enables detection directly link the segmentation tasks to consider the contribution of background information for optimization. It provides a better optimization of both DRL and GAN in the context of semi-supervised learning.
- A bidirectional exploration mechanism combines forward and backward exploration direction, enabling the agent to be guided in the correct direction when the exploration is hindered due to a lack of accurate rewards.

II. Related work

A. Semi-supervised medical image segmentation methods

Various direct mapping based semi-supervised medical image segmentation methods have been published currently, but there is no DRL-based semi-supervised method. These methods [11], [42] can be summarized as following four categories: 1) consistency regularization-based methods [5], [31], [32], [40], which assume that model should have a consistent prediction on a given unlabeled example; 2) pseudo-label-based methods [41], [44], which leverage a trained model on the labeled set to produce additional training examples by labeling unlabeled data; 3) generative model-based methods [42], which learn the data distribution from labeled and unlabeled data to generate more training data; 4) graph-based methods [43], which consider labeled and unlabeled data as nodes of a graph, thus enabling label information to propagate from labeled data to unlabeled data. These methods perform satisfactorily on most tasks. However, all the above direct mapping based methods seek the global optimal among the whole image. In this case, these methods are often limited in the segmentation of small or low-contrast objects because they make weak contributions to global optimization. As mentioned in the introduction, DRL's sequential iterative exploration mechanism enables it to address this problem. Our method represents a pioneering effort in the development of a DRL-based semi-supervised medical image segmentation method.

B. DRL methods with reward not directly calculated from labels

The reward function serves as a crucial link between the task goal of the DRL method and the action taken. A well-designed reward function should accurately reflect the task goal and incentivize the DRL to take the appropriate action [2], [20]. There are various mainstream reward function design methods, such as directly using labels to calculate distance and IoU between the virtual box where the agent moves and the bounding box defined by the label as rewards, or employing Inverse Reinforcement Learning to infer reward functions from expert demonstrations [12], [13]. However, these reward functions have limitations, particularly in supporting semi-supervised training. Recent research has explored the use of an auxiliary discriminator to provide reward signals is a promising solution, which trains a discriminator network to evaluate DRL actions. This method have been apply for tasks such as label refinement and pseudo-label generation [27], but these methods still have problems such as still relying on other label based rewards or cannot be used for end-to-end training. Our approach creates a stable, mutually complementary reward function by using GAN to perform segmentation with the DRL detection and simultaneously evaluate the detection accuracy. The detection and segmentation complement each other and form a stable reward that supports end-to-end training.

C. Prior-based exploration

Exploration is an important issue in reinforcement learning, as the effectiveness of reinforcement learning methods is directly determined by the effect of exploration [24], [26]. There exists a variety of exploration strategies [21], [28], [29], such as the $\epsilon-$greedy strategy [28], which randomly selects an action with a probability of ϵ and chooses the known optimal action with a probability of $1-\epsilon$. The Upper Confidence Bound strategy [29] calculates the upper confidence bound of each action and chooses the action with the largest confidence. Maximum entropy exploration [21] maximizes the entropy of the strategy to achieve relatively uniform probabilities for all actions. However, these strategies often perform poorly in sparse reward environments, such as those where rewards are provided by GAN. This is because GAN cannot directly provide rewards during the initial exploration stage of DRL. Sparse rewards hinder DRL from selecting optimal actions, rendering stochastic exploration methods less effective. Recent research has proposed backward exploration as a potential

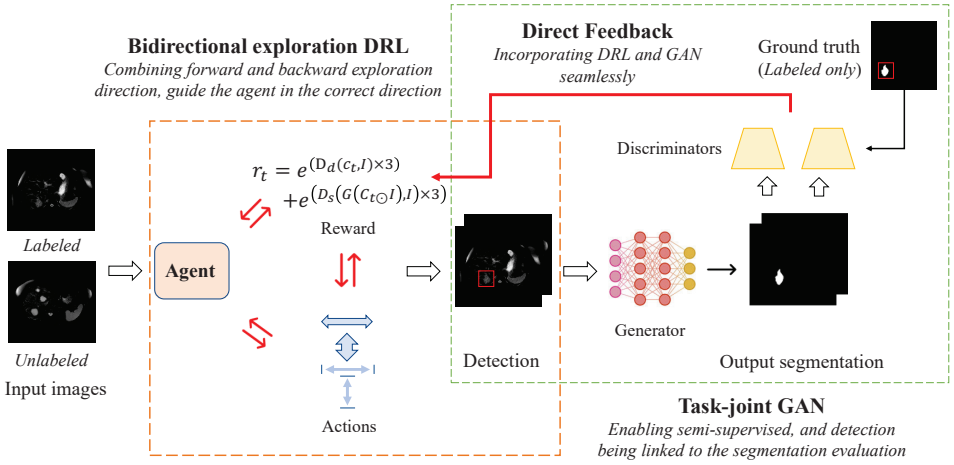

Fig. 3: DGARL ingeniously integrates bidirectional exploration DRL and Task-Joint GAN. The bidirectional exploration DRL employs both forward and backward exploration to facilitate accurate detection from input images. Concurrently, the Task-Joint GAN accomplishes semi-supervised segmentation, executing a joint evaluation of both segmentation and detection tasks. The assessment outcomes from the GAN are reciprocated back to the DRL, guiding its modifications for improved segmentation region production. Consequently, the heightened accuracy of the DRL further propels the GAN to generate superior segmentations.

solution to the sparse rewards problem [14], [15]. With backward exploration, reinforcement learning explores from the completed state back to the previous state, providing prior experiences of successfully completing the task and helping the agent learn how to make optimal choices in complex environments. This method has been applied to help robots perform action planning in complex environments, but it has yet to be used in medical image analysis tasks. Our approach utilizes bidirectional exploration consisting of forward exploration and backward exploration to gain successful prior experiences to solve the sparse reward problem introduced by GAN.

D. Existing RL + GAN methods

Although there are some RL+GAN methods applied in other fields, the current no RL+GAN methods applied for image segmentation due to the drawbacks of GAN's rewards as described in the introduction. Existing RL+GAN methods can be divided into two categories based on the relationship between RL and GAN: 1) independent RL and GAN [35], [36], where RL and GAN run independently without direct interaction, and RL is trained based on the data generated or completed by GAN. For example, in [35], GAN is used to simulate an environment and RL controls a robot in the synthetic environment to grasp real-world objects. 2) RL as part of the GAN generator [37]–[39], where RL is integrated into the GAN generator and its reward comes from GAN. GAN's discriminator evaluates the difference between RL actions and expert actions, or between generated results and target results. For instance, [38] uses the discriminator to distinguish between states generated by RL and those generated by an expert to encourage RL to imitate expert strategies. Additionally, [39] uses RL to control a pen for drawing, and GAN evaluates the difference between the current drawing board and the target image to encourage RL to create human-level art. Our method integrates RL into the generator, utilizes RL for detection, performs fine-grained segmentation based on RL-based detection, and directly uses the evaluation of the GAN that combines segmentation and detection as the reward for RL. Furthermore, we employ a novel bidirectional exploration strategy to overcome the drawbacks of GAN rewards in RL field.

III. Methodology

As a DRL-based semi-supervised segmentation method, as shown in Fig. 3, DGARL inputs both labeled and unlabeled images, iteratively explores them, performs segmentation based on the exploration, and ultimately outputs accurate segmentation results. It integrates Bidirectional Exploration DRL and Stage-confluent GAN. The Bidirectional Exploration DRL learns from forward-backward combined exploration experiences to estimate the most appropriate region for medical object segmentation. The Stage-confluent GAN not only performs segmentation based on the estimated region but also evaluates segmentation and detection to provide accurate feedback. The region estimated by the Bidirectional Exploration DRL prevents the Stage-confluent GAN from being distracted by the background, while the Stage-confluent GAN provides rewards for the Bidirectional Exploration DRL to estimate the appropriate region. In other words, the Bidirectional Exploration DRL and the Stage-confluent GAN mutually promote each other to achieve accurate segmentation. DGARL learns to enhance segmentation quality on unlabeled data to obtain higher evaluations from the discriminator, thereby enabling semi-supervised training.

A. Bidirectional exploration DRL

The bidirectional exploration DRL explores images from forward and backward, respectively. Backward exploration provides experience with dense rewards which enlarged the state space for DRL to explore and complement the sparse rewards problem from forward exploration, thus providing

accurate object detection. The bidirectional exploration DRL inputs an image and outputs the most appropriate region to perform segmentation by iteratively exploring the image forward and backward, respectively, thereby assisting the Stage-confluent GAN in getting an accurate mask for the desired medical object. In the rest of this section, we will firstly introduce the Forward-backward combined exploration. Subsequently, we will describe the elements in the DRL for the iterative estimation process.

1) Forward-backward combined exploration: Forward-backward combined exploration introduces backward exploration to complement forward exploration, providing prior experiences that effectively alleviate the sparse reward problem and facilitate the convergence of the DRL. Forward-backward combined exploration provides additional experience to guide DRL to reach the target smoothly from the target, greatly relieving the exploration pressure caused by the lack of guidance of the intermediate reward in single forward exploration [15]. In the Forward-backward combined exploration, the forward exploration is from the image center to the target, and the backward exploration is from the target to the surroundings which is stochastic and prior. Specifically, in forward exploration, at each time t, the agent observes the environment state s_t and selects an action a_t based on current policy π_θ. DRL executes a_t to change the state and enters the next time step $t+1$, receiving a numerical reward r_{t+1} and a new state s_{t+1}. Thus, the trajectory $\tau_{forward}$ generated by the agent's iteration in the environment can be expressed as :

$$\tau_{forward} = (s_0, a_0, r_1, s_1, a_1, r_2, s_2, a_2, r_3, ...) \tag{1}$$

Backward exploration starts from the target and explores the trajectory backwards, denoting the artificially generated state of the complete detection of the target as s_i. the backward exploration randomly generates an action a_{i-1}, and the DRL performs the inverse action $-a_{i-1}$ to get the previous state s_{i-1} and calculate the corresponding reward $r_i = R(s_{i-1}, a_{i-1})$. By repeating this process and back-warding the exploration experience, the bidirectional exploration DRL can obtain trajectory $\tau_{backward}$:

$$\tau_{backward} = (..., r_{i-1}, s_{i-1}, a_{i-1}, r_i, s_i) \tag{2}$$

Such backward exploration provides the experience of successfully detecting the target. The DRL can learn from the experience of backward exploration how to detect the target and the meaningful background around the target.

2) Elements for iterative estimation process: The bidirectional exploration DRL initializes a virtual box on the input image to indicate the optimal area for segmentation. By observing the virtual box as the state s, the bidirectional exploration DRL trains to select an action a and refine the virtual box step-by-step. The bidirectional exploration DRL sends the virtual box to the Stage-confluent GAN and receives feedback evaluation to distribute reward r. The bidirectional exploration DRL pursues the maximization of rewards, and will guide the virtual box to approach the desired object and terminate at the best perspective for segmentation.

State. The state s_t at iteration t is defined by a tuple $[\mathcal{C}_t, \mathcal{I}]$, where $\mathcal{I}$ represents the original input image. $\mathcal{C}_t$ represents a binary map of the same size as $\mathcal{I}$, where the pixel intensity inside the virtual box is one while outside the box is zero.

Action. The action a_t selected at time t for state s_t is defined by a tuple $[a_t^{tv}, a_t^{th}, a_t^{vs}, a_t^{hs}]$ to refine the box flexibly. a_t^{tv} and a_t^{th} represent the magnitude of vertical and horizontal movement of the box, respectively. a_t^{vs} and a_t^{hs} represent the magnitude of the change in the length and width of the box, respectively.

Reward function. The reward function is determined by the Stage-confluent GAN, which comprehensively evaluates the region estimation and the resulting segmentation. The reward function can be formulated as:

$$r_t = \mathrm{Score}(\mathcal{I}, \mathcal{C}_t) \tag{3}$$

Where Score is the output of the Stage-confluent GAN after inputting the image $\mathcal{I}$ and the box-built map $\mathcal{C}_t$. This reward function facilitates the agent to approach the desired object and seek the best perspective for segmentation simultaneously.

Agent. The agent follows the Soft Actor-Critic (SAC) [21] structure, which is a maximum entropy reinforcement learning algorithm that has stronger exploration capabilities and performs more stably than other algorithms in complex environment.

B. Stage-confluent GAN

The Stage-confluent GAN performs segmentation on the region estimated by the bidirectional exploration DRL and utilizes dual discriminators to evaluate the segmentation performance and the effectiveness of the estimated region, respectively, thus to provide accurate segmentation and overall evaluation for both detection and segmentation. In the Stage-confluent GAN, the segmentation evaluation is fed back to the internal generator for semi-supervised segmentation training. Meanwhile, the overall evaluation is fed back to the bidirectional exploration DRL and built as the reward, which guides the bidirectional exploration DRL to offer the best region for the Stage-confluent GAN. In the rest of this section, we will firstly introduce the architecture of the Stage-confluent GAN. Subsequently, we will describe the loss function training the Stage-confluent GAN.

1) GAN architecture with dual discriminators for comprehensive evaluation: The Stage-confluent GAN consists of a generator G, a segmentation discriminator D_s, and a region discriminator D_d. The generator predicts the desired object mask. The dual discriminators provide supervision for the generator to achieve semi-supervised training and provide the evaluation Score for the bidirectional exploration DRL.

The generator G builds a binary map $\mathcal{C}$ with the box outputted by the bidirectional exploration DRL, thus obtaining the dot product $\mathcal{I} \odot \mathcal{C}$. Thus, the generator successfully avoids the segmentation from the unrelated background information. By inputting the above dot product to a ResUNet generator, the generator gets a segmentation mask M. Here, the ResUNet can be replaced with other segmentation networks that fit the given task. More specifically, because our data size is

small (hundreds of cases) and our task is relatively simple (most unrelated background has been removed by the bidirectional exploration DRL), we selects the U-Net to validate the performance of our framework.

The segmentation discriminator and the detection discriminator have the same structure, which consists seven convolution layers to decrease the feature dimension, and four fully-connected layers to decide the performance of the segmentation or the region estimation. D_s takes the tuple $[\mathcal{I}, M]$ as a fake sample and tuple $[\mathcal{I}, y]$ as a real sample and outputs the evaluation of segmentation, where y is the manually-annotated object mask. D_d takes the tuple $[\mathcal{I}, \mathcal{C}_t]$ as a fake sample and tuple $[\mathcal{I}, \mathcal{B}]$ as a real sample, outputs the evaluation of region estimation, where $\mathcal{B}$ is the binary map built with the manually-annotated bounding-box. Therefore, the segmentation is closer to the manual annotation, and the output of D_s is closer to 1, which also applies to D_d. By combining the above two evaluations, the score for the bidirectional exploration DRL is:

$$\text{Score} = (D_d([\mathcal{C}, \mathcal{I}]) + D_s([G(\mathcal{C} \odot \mathcal{I}), \mathcal{I}])) \times \alpha \qquad (4)$$

The term $D_d([\mathcal{C}, \mathcal{I}])$ is to guide the bidirectional exploration DRL to approach the desired object, the term $D_s([G(\mathcal{C} \odot \mathcal{I}), \mathcal{I}])$ is to guide the bidirectional exploration DRL to find the perspective beneficial to the segmentation, and the hyper-parameter α is a variable that controls the magnitude of the reward and is empirically set to 3.

2) Adversarial loss function for semi-supervised training: The Stage-confluent GAN supports semi-supervised training, where the generator and discriminator have different losses for the labeled set $\{\mathcal{I}^l, y^l\}$ and unlabeled set $\{\mathcal{I}^u\}$. At every time step t, the box outputted by the bidirectional exploration DRL is built as a binary mask $\mathcal{C}_t$. The loss training the generator is:

$$\mathcal{L}_G = \begin{cases} \mathcal{L}_T(\mathcal{I}^l, \mathcal{C}_t, y^l) + \mathcal{L}_{\text{BCE}}(\mathcal{I}^l, \mathcal{C}_t, 1) & \text{for labeled cases} \\ \mathcal{L}_{\text{BCE}}(\mathcal{I}^u, \mathcal{C}_t, 1) & \text{for unlabeled cases} \end{cases}$$

Where $\mathcal{L}_T$ is Tversky loss [22]:

$$\mathcal{L}_T(\mathcal{I}, \mathcal{C}, y) = \mathbf{Tversky}(G(\mathcal{I} \odot C), y)$$

The Tversky loss increases the weight of false positive and false negative samples, thus balancing the precision and recall of medical image segmentation. $\mathcal{L}_{\text{BCE}}$ is the Binary Cross Entropy(BCE) loss, which leverages the segmentation discriminator to achieve semi-supervised learning:

$$\mathcal{L}_{\text{BCE}}(\mathcal{I}, \mathcal{C}, d) = \mathcal{L}_{\text{BCE}}(D_s(G(\mathcal{I} \odot C), \mathcal{I}), d)$$

Where d is the pseudo label. The loss functions to train two discriminators are:

$$\mathcal{L}_{D_s} = \mathcal{L}_{\text{BCE}}(\mathcal{I}^l, \mathcal{C}_t, 0) + \mathcal{L}_{\text{BCE}}(D_s(\mathcal{I}^l, y^l), 1)$$

$$\mathcal{L}_{D_d} = \mathcal{L}_{\text{BCE}}(D_d(x_l, \mathcal{C}_t), 0) + \mathcal{L}_{\text{BCE}}(D_d(x_l, C_l), 1)$$

IV. Experiment

A. Data acquirement

Brain tumor dataset [45], [46]: The brain tumor dataset was acquired from Nanfang Hospital, Guangzhou, China, and General Hospital, Tianjing Medical University, China, from 2005 to 2010. The brain tumor dataset containing 3064 T1-weighted contrast-enhanced images from 233 patients with three kinds of brain tumors: meningioma (708 slices), glioma (1426 slices), and pituitary tumor (930 slices). We used the first 1008 images out of the 3064 images to speed up verification. The brain tumor dataset is a low-contrast dataset, as brain MRI is widely recognized for its noisy, low-contrast features [49], [50] , and the contrast between tumor tissue and gray matter is low [51], [52].

Pancreas-NIH dataset [30]: The Pancreas-NIH dataset consists of 82 abdominal contrast enhanced 3D CT scans (70 seconds after intravenous contrast injection in portal-venous) from 53 male and 27 female subjects. The CT scans have resolutions of 512x512 pixels with varying pixel sizes and slice thickness between 1.5 - 2.5 mm, acquired on Philips and Siemens MDCT scanners (120 kVp tube voltage). The Pancreas-NIH dataset is both a low-contrast and small-objects dataset, given the indistinct boundaries of the pancreas [11,12] and an average target-to-image ratio of 0.0041, which is below the recognized standard of 0.01 for small objects [53]–[55].

Liver tumor dataset [1], [47], [48]: The liver tumor dataset consists of 325 liver tumor cases including benign tumors (hemangiomas, 100), malignant tumors (hepatocellular carcinoma, 150), and normal controls (75). Contrast-enhanced and non-enhanced liver MRI images were obtained using 1.5 T MRI scanners (two types: Signa Artist, GE, and Aera, Siemens) using T1-weighted imaging (0.1 mmol/kg dose of gadolinium-based CAs). The liver tumor dataset is considered a small-objects dataset, with an average target-to-image ratio of 0.0056, encompassing a significant number of small objects. This ratio is much smaller than the recognized standard of 0.01 for small objects [54]–[58].

B. Hyper-parameter and Environments.

All hyperparameters in DGARL have been verified or fine-tuned to be optimal, and they were fine-tuned using only the training set. The hyperparameters in DGARL can be divided into basic hyperparameters and key hyperparameters. The basic hyperparameters of the DRL part include forward exploration step size, action scale and network structure, etc., and they all follow the settings of [1]. Learning rates of the DRL and GAN's generator start with 1e-4, and the learning rate of the discriminator of GAN starts from 1e-5, and gradually decrease based on the decay coefficient of 0.95. The key hyperparameters in DGARL include the type of DRL method as SAC, the backward exploration step size as 30, and the proportion of backward exploration experience usage as 50%, which are verified as optimal by hyperparameter selection experiments. All the image is resized to 128×128. DGARL is implemented with PyTorch 1.8.0 on 1 NVIDIA RTX 3090 and Intel Xeon Gold 6242R under Ubuntu 20.08.

C. Training process

Algorithm 1 Main training process

Input: Labeled image set $\{X_L, Y_L\}$ and unlabeled image set $\{X_U\}$
Output: Segmentation masks of input

initialize experience replay memory $\mathrm{Mem}_{forward}$ and $\mathrm{Mem}_{backward}$
for image $\mathcal{I}$ in $\{X_L, X_U\}$ **do**
  Forward exploration
  initialize first state $s_t(t=0)$
  while $t < L$ **do**
    action a_t is sampled from policy π
    take action a_t, receive next state s_{t+1}
    input state s_{t+1} into GAN to perform segmentation
    compute reward r_{t+1} based on the Stage-confluent GAN evaluation
    update the Stage-confluent GAN
    add tuple$(s_t, a_t, s_{t+1}, r_{t+1})$ to $\mathrm{Mem}_{forward}$
    $t++$
  end while
  if image i belongs to $\{X_L\}$ **then**
    Backward exploration
    initialize final state $s_t(t=L)$
    while $t > (L-30)$ **do**
      action a_{t-1} is sampled randomly
      calculate previous stable s_{t-1} of execute a_{t-1}
      input state s_t into the Stage-confluent GAN to perform segmentation
      compute reward r_t based on the Stage-confluent GAN evaluation
      update the Stage-confluent GAN
      add tuple$(s_{t-1}, a_{t-1}, s_t, r_t)$ to $\mathrm{Mem}_{backward}$
      Sample random batch from $\mathrm{Mem}_{forward}$ and $\mathrm{Mem}_{backward}$
      Update the bidirectional exploration DRL
      $t--$
    end while
  end if
end for

D. Experimental setting

We conducted inter-comparison experiments, ablation experiments, and hyper-parameter selection experiments to comprehensively validate our proposed framework. All experiments were conducted based on the same data partition, and the performance is measured by 6 metrics including dice coefficient (Dice), 95% Hausdorff distance (95HD), intersection over union (IoU), Cohen Kappa coefficient (Kappa), Sensitivity and Specificity [34].

1) Inter-comparison experiments: Inter-comparison experiments evaluate the performance of DGARL compared with baseline and state-of-art methods. Particularly, the inter-comparison experiments include:

a) Baseline methods: Uncertainty-aware Mean-Teacher (UAMT) [5], Mask-RCNN [6], UNet [33] , Deep q-network-driven catheter segmentation (DQN) [3], Multi-step medical image segmentation (MSMIS) [4]

b) State-of-art methods: Bidirectional copy-paste for semi-supervised medical image segmentation (BCP) [32], Exploring smoothness and class-separation for semi-supervised medical image segmentation (SSNet) [31], Boundary mining with adversarial learning (BMA) [8] , Reinforced active learning for image segmentation (RALIS) [7], Weakly-supervised teacher-student (WSTS) [1].

First, we used 4 different training/test data ratios (20:80, 40:60, 60:40, 80:20) for experiments, in which 50% of the training data was set as labeled data, and Supervised methods (DGARL, BCP, BMA, SSNet and UAMT) use an additional 50% of unlabeled data.

Secondly, we also compared our method to BCP, BMA, SSNet, UAMT and Mask-RCNN under 80:20 training/test data split using 25% and 75% labeled data respectively to test the robustness of our method.

Finally, we conduct incremental comparative experiments using pretrained models on the unseen Pancreas-NIH dataset with different modalities to test the generalization ability of our method and to avoid potential data snooping issues. Specifically, we use semi-supervised methods (DGARL, BCP, BMA, SSNet, UAMT) trained under the brain tumor MRI dataset (80:20 training/test data ratio) as pre-trained models. Increasing training data (0%, 10%, 20%, 30%) were used for comparison on the new Pancreas-NIH dataset, where 50% in the training data was set as labeled data.

2) Ablation experiments: Ablation experiments evaluate the effect of our proposed Bidirectional exploration DRL and Stage-confluent GAN by eliminating key components in our framework.

The ablation experiment is set as: based on DGARL framework, 1)employing single segmentation discriminator reward and single detection discriminator reward; 2)employing forward exploration only and abandon backward exploration.

3) Hyper-parameter selection experiments: The hyper-parameter selection experiments evaluate the setting of those key hyper-parameters by controlling the proportion of forward and backward exploration experience used in DRL update, the step length of backward exploration and the selection of DRL method. Specifically, the hyper-parameter selection experiment includes: 1) Set the proportion of backward exploration experience to 50%, and use backward exploration steps of 30, 20 and 10, respectively. 2) Set the step size of backward exploration to 30, and use 70% and 30% backward exploration experience ratios respectively. 3) Set the proportion of backward exploration experience to 50%, and the step size to 30, and use SAC, DDPG [9] and TD3 [10] as the core DRL methods respectively.

E. Experimental results

Comprehensive experiments have demonstrated that DGARL is capable of generating accurate image segmentation results. In particular, under 80:20 training/test data ratios, DGARL achieves a Dice score of 85.02%, a Hausdorff distance of 4.64 at 95% confidence level, an IoU of 77.51%

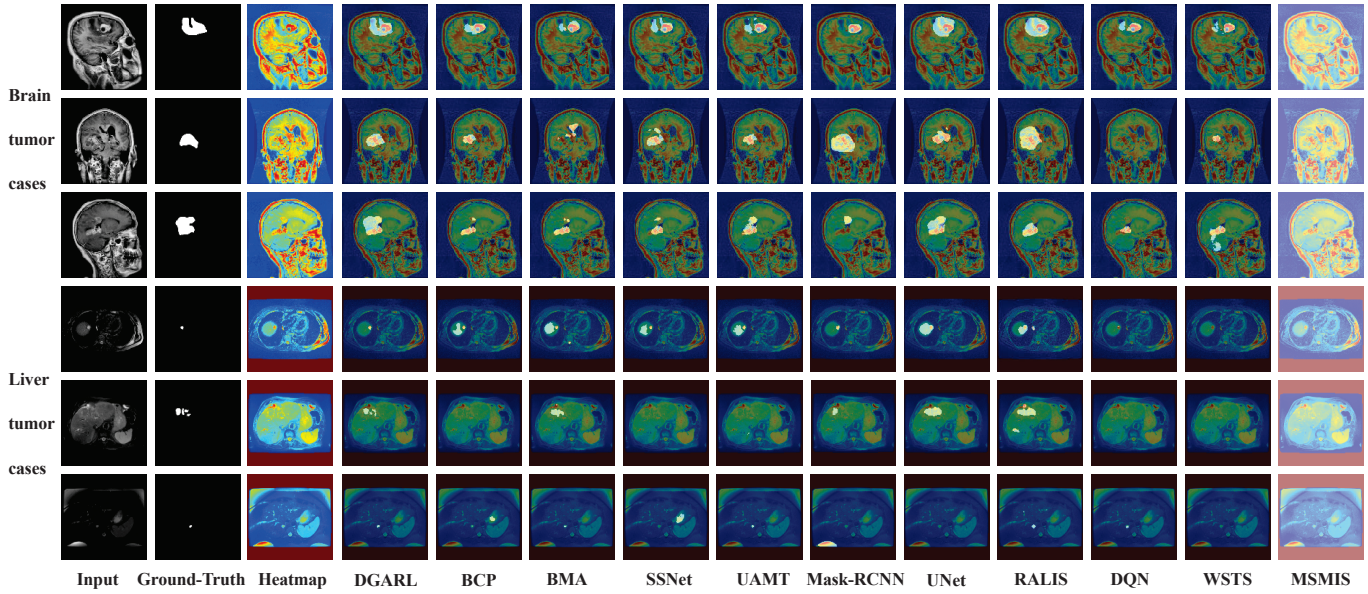

Fig. 4: Visual comparisons between DGARL and other methods under 80:20 training/test data ratios. The first column shows the input images. The second column shows the ground-truth. The third column shows the input image converted to a heatmap to highlight details. The remaining columns illustrate the segmentation of the eleven methods on the input image, and the results are highlighted on the heatmap.

and a Kappa of 84.75% in brain tumor dataset. Similarly, DGARL achieves a Dice score of 73.18%, a Hausdorff distance of 3.55 at 95% confidence level, an IoU of 66.58%, a Kappa of 72.82% in liver tumor dataset. These results demonstrate that DGARL has great potential to serve as an accurate automatic segmentation method for medical images with low-contrast or small objects.

1) The highest segmentation accuracy of DGARL shows through Inter-comparison experiments: TableI and Fig 4 indicate that DGARL achieves accurate lesion segmentation. Specifically, compared with SOTA methods, under 80:20 training/test data ratios, in the brain tumor dataset, DGARL improved the Dice coefficient by 1.92%. And in the liver tumor dataset, DGARL improved the Dice coefficient by 4.28%. These results demonstrate that the segmentation of DGARL are more accurate and have higher consistency with annotations than other methods. More intuitively, fig 4 visually shows that compared with other segmentation methods. From fig 4, we can see that DGARL improves the accuracy of image segmentation and avoids mis-segmentation on background regions that similar to the object. All these improvements are because DGARL uses DRL to limit the area of segmentation

TABLE I: DGARL produced higher quality segmentation results than existing segmentation methods under 80:20 training/test data ratios. All of the methods have employed 50% of labeled data, while semi-supervised methods (DGARL, BCP, BMA, SSNet, UAMT) have further utilized 50% of unlabeled data.

Dataset	Method	Metrics					
		Dice(%)	95HD	IOU(%)	Kap(%)	Sensitivity(%)	Specificity(%)
Brain-tumor	DGARL	**85.02**	**4.64**	**77.51**	**84.75**	87.72	99.72
	BCP	80.79	5.05	72.62	80.47	79.84	99.83
	BMA	78.64	5.88	68.61	78.21	86.96	99.49
	SSNet	83.11	4.65	73.33	82.83	81.95	99.86
	UAMT	78.19	5.93	75.58	77.81	77.68	99.81
	Mask RCNN	79.84	6.08	69.21	79.56	82.03	99.65
	UNet	75.78	5.77	63.79	75.33	89.05	99.42
	RALIS	78.89	5.37	68.17	78.48	92.51	99.41
	DQN	65.75	7.03	57.06	65.29	62.87	99.85
	MIMIS	3.60	121.6	1.85	0	100	0
	WSTS	66.73	6.96	58.73	66.30	63.34	99.89
Liver-tumor	DGARL	**73.18**	**3.55**	**66.58**	**72.82**	70.96	99.87
	BCP	63.58	4.43	55.48	62.96	62.13	99.92
	BMA	68.90	4.06	60.23	68.41	68.76	99.89
	SSNet	68.06	4.03	61.19	67.53	69.02	99.90
	UAMT	66.25	4.11	58.99	65.64	66.95	99.89
	Mask RCNN	63.07	3.62	56.23	62.73	66.78	87.64
	UNet	58.54	5.20	47.49	57.86	73.21	99.70
	RALIS	63.63	5.58	51.19	63.20	85.89	99.66
	DQN	63.77	4.30	55.00	63.04	59.67	99.96
	MIMIS	1.08	121.6	0.54	0	100	0
	WSTS	63.76	4.15	56.31	62.94	59.59	99.96

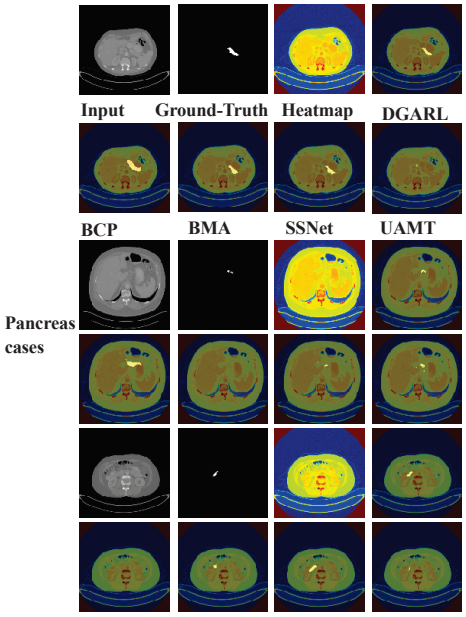

Fig. 5: Visual comparisons between DGARL and other methods on Pancreas-NIH dataset under using 30% extra training data, which demonstrates the better generalization ability of DGARL than other methods.

TABLE II: DGARL performs stable and more accurate segmentation than other methods on different training/test data ratios.

Dataset	Method	Dice(%) under different data division			
		20:80	40:60	60:40	80:20
Brain Tumor	DGARL	**71.80**	**75.97**	**79.36**	**85.02**
	BCP	63.90	72.73	77.27	80.79
	BMA	63.66	74.68	77.90	78.64
	SSNet	69.65	73.71	77.52	83.11
	UAMT	66.02	72.28	75.01	78.19
	Mask RCNN	67.32	69.17	75.07	75.51
	UNet	65.19	71.78	72.35	75.78
	RALIS	60.92	70.75	74.06	78.89
	DQN	53.22	48.61	62.47	65.75
	WSTS	50.51	53.12	58.43	66.73
	MSMIS	3.58	3.62	3.75	3.60
Liver Tumor	DGARL	**57.04**	**65.02**	**68.17**	**73.18**
	BCP	41.79	52.40	57.30	63.58
	BMA	54.07	57.26	63.22	68.90
	SSNet	54.08	62.66	65.77	68.06
	UAMT	42.96	51.92	65.19	66.25
	Mask RCNN	52.59	59.63	62.77	63.07
	UNet	40.01	48.03	53.83	58.54
	RALIS	32.14	49.06	60.75	63.63
	DQN	31.00	29.79	49.58	63.77
	WSTS	34.16	43.52	60.05	63.76
	MSMIS	1.13	1.09	1.12	1.08

which simplifies segmentation tasks avoid mis-segmentation caused by complex backgrounds. Moreover, DGARL selects the area to segment based on the performance of the segmentation network, which retains background information that is beneficial to segmentation, and meanwhile avoids ignoring edge information. Furthermore, compared to methods solely utilizing detection as an auxiliary task (e.g., DQN, WSTS) and those adopting a multi-task approach combining detection and segmentation (e.g., MaskRCNN), our method demonstrates an improvement in the Dice coefficient by 5.18%-9.41%. Compared with these methods, our method prevents error accumulation, resulting in more accurate segmentation.

Table II presents the superior Dice scores achieved by DGARL compared to other methods at four different training/testing data ratios (20:80, 40:60, 60:40, 80:20). DGARL outperforms other methods by at least 2.15%-2.96%, 1.29%-2.36%, 1.46%-2.4%, and 1.91%-4.28% in Dice score respectively. Table III demonstrates the superior Dice and IoU scores achieved by DGARL compared to other methods at two different labeled data ratios (25%, 75%). Under 25% labeled data, DGARL achieves at least 2.59%-2.66% higher Dice score and 1.54%-3.31% higher IoU score than other methods. Under 75% labeled data, DGARL achieves at least 1.21%-3.71% higher Dice scores and 2.73%-3.29% higher IoU scores than other methods. These results indicate the robustness of our approach in handling diverse data scales.

Table IV and Figure 5 present the incremental comparison results of our method against other semi-supervised methods on the Pancreas-NIH dataset. Specifically, our method outperforms the others in terms of Dice score by 3.85% with an additional 10% data, by 1.02% with an additional 20% data, and by 2.73% with an additional 30% data. These results demonstrate the ability of our method to effectively generalize to new data domains with limited additional data.

2) Superiority of the bidirectional exploration DRL: Fig 6 demonstrates that bidirectional exploration in our framework is crucial for achieving accurate detection and segmentation in our tasks. The use of bidirectional exploration leads to significant improvements in segmentation metrics. These improvements are attributed to the fact that the reward used by DRL is computed by GAN, which cannot provide direct rewards for the early stage of DRL exploration. Bidirectional exploration DRL can still provide accurate experience for DRL through backward exploration even when forward exploration is unable to provide reliable experience. Bidirectional exploration ensures that DRL can achieve accurate detection in complex and reward-sparse environments.

3) Advantage of the GAN based detection and segmentation reward: Fig 6 demonstrate that the use of detection and segmentation rewards can enhance the performance of DGARL. Detection reward is simpler and easier to obtain than segmentation reward, and it can accelerate the convergence of DGARL during training. On the other hand, segmentation reward directly reflects the quality of the segmentation result, which motivates DRL to select the most suitable segmentation area. However, obtaining the segmentation reward is more challenging than obtaining the detection reward, and the generator requires more iterations to generate high-quality segmentation results. DGARL using the combination of segmentation reward and detection reward can quickly learn and achieve higher-precision detection. Therefore, the use of combination of segmentation reward and detection reward in DGARL can significantly improve the performance of the model and provide more accurate segmentation results.

4) Optimal training parameters chooses through hyper-parameter selection experiments: Table III presents the performance of DGARL under different settings of backward exploration step length and proportion of backward exploration

TABLE III: DGARL maintains outperforms other methods under different labeled/unlabeled data ratios

Dataset	Method	25% labeled data		75% labeled dat	
		Dice(%)	IoU(%)	Dice(%)	IoU(%)
Brain Tumor	DGARL	**77.23**	**69.51**	**85.96**	**79.31**
	BCP	74.57	66.20	81.75	73.04
	BMA	72.15	62.24	80.94	71.93
	SSNet	71.81	63.20	84.75	76.58
	UAMT	68.91	60.01	78.08	69.66
	Mask RCNN	69.68	62.81	83.17	75.98
Liver Tumor	DGARL	**59.03**	**50.09**	**80.45**	**66.58**
	BCP	54.01	45.39	64.53	55.48
	BMA	55.18	46.08	76.74	60.23
	SSNet	53.38	45.97	75.62	61.19
	UAMT	56.44	48.48	72.53	58.99
	Mask RCNN	55.05	48.55	71.08	63.29

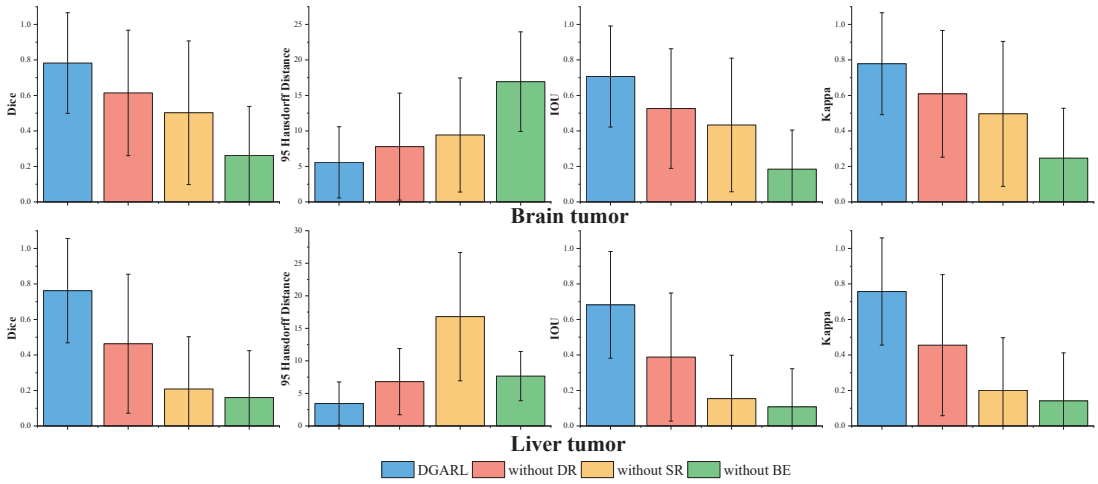

Fig. 6: Performance comparison of the DGARL framework with and without the use of detection reward (DR), segmentation reward (SR), and bidirectional exploration (BE). The joint use of these components resulted in higher Dice, IOU, and Kappa scores (higher is better) and lower 95HD scores (lower is better), highlighting the effectiveness of our proposed novel components.

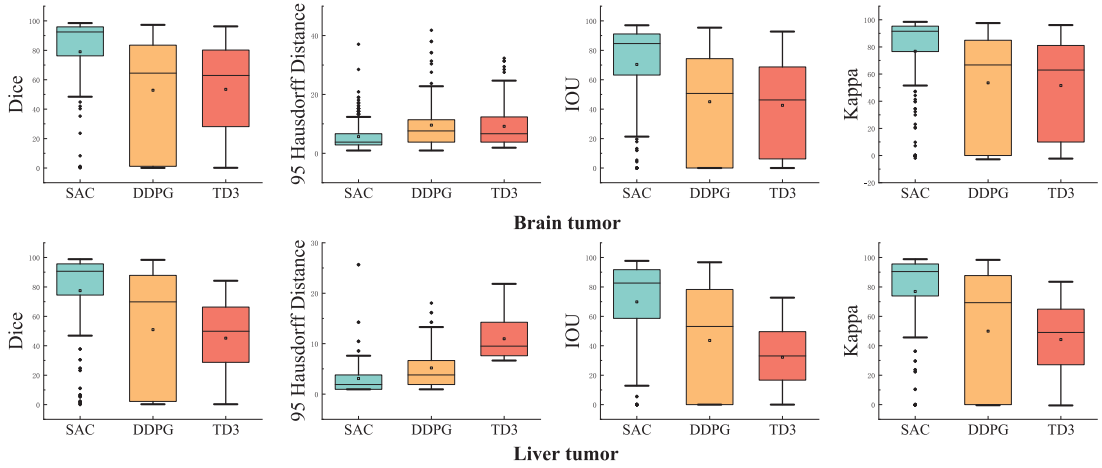

Fig. 7: DGARL using SAC as core DRL part achieves higher performance than using DDPG and TD3.

TABLE IV: DGARL maintains outperforms other methods under different extra data.

Methods	Dice(%) under different extra data			
	0%	10%	20%	30%
DGARL	1.51	**54.67**	**64.36**	**70.85**
BCP	1.03	2.4	22.24	30.60
BMA	2.94	50.82	62.94	67.43
SSNet	1.10	1.10	63.34	68.12
UAMT	0.46	32.07	44.24	51.43

experience used in the DRL update. The results show that the best performance is achieved when the backward exploration step length is set at 30 and the proportion of backward exploration experience used in the DRL update is set at 50%. And the experiment results suggest that a longer backward exploration step length can improve the segmentation performance in both tasks by providing more background information around the target. However, excessively increasing

TABLE V: DGARL produced the best performance by tuning the main network parameters, including the backward step length and the proportion of forward and backward exploration experience used in DRL update.

Dataset	backward step length	Percentage of backward experience used	Metrics			
			Dice(%)	95HD	IoU(%)	Kappa(%)
Brain-tumor	30	50%	**78.24**	5.57	**70.68**	**77.87**
	20	50%	77.65	**5.46**	69.60	77.27
	10	50%	53.59	9.19	45.09	52.92
	30	70%	63.75	7.64	56.09	63.23
	30	30%	61.30	8.01	52.88	60.75
Liver-tumor	30	50%	**76.25**	**3.45**	**68.21**	**75.73**
	20	50%	62.94	4.70	55.24	62.25
	10	50%	53.70	5.39	46.19	52.81
	30	70%	63.24	4.73	55.72	62.66
	30	30%	60.30	4.69	52.53	59.59

the backward step length does not bring significant improvements, and taking the same length as forward exploration is sufficient. For instance, increasing the backward step length from 20 to 30 in brain tumor segmentation only increased dice by 0.59%. Moreover, the performance reaches its best when the proportion of backward exploration experience used in the DRL update is set at 50%. This is because backward experience comes from stochastic actions, while the ultimate goal of SAC is to optimize actions from the current policy π. Therefore, an excessive proportion of backward experience will lead to neglect of policy π, while a low proportion is unable to provide enough guidance.

Fig 7 indicates that DGARL using SAC outperforms DGARL using DDPG and TD3. Specifically, DGARL using SAC showed improvements in multiple metrics such as Dice, 95HD, IoU, and Kappa. These results indicate that DGARL using SAC achieves higher performance and lower degree of dispersion, making it more accurate and stable than DDPG and TD3. The superior performance of SAC can be attributed to its use of an entropy maximization strategy, which ensures that potentially valuable actions are not missed. This makes SAC more robust than DDPG and TD3 in complex environments. Although DDPG and TD3 use noised actions to improve their stochastic exploration capabilities, they still lack the ability to explore all possible optimal paths.

TABLE VI: The model size of each method.

Methods	Model Size in training (params)	Model Size in prediction (params)
DGARL	36.21 M	9.55 M
BCP	3.24 M	3.24 M
BMA	22.33 M	12.75 M
SSNet	3.24 M	3.24 M
UAMT	3.24 M	3.24 M
Mask RCNN	43.92 M	43.92 M
UNet	3.24 M	3.24 M
RALIS	3.6 M	3.24 M
DQN	9.5 M	9.5M
MIMIS	18.32 M	6.16 M
WSTS	22.07 M	9.55 M

F. Implication on the size of the model

Table VI presents the model size of each method in the compared experiments. During the training phase, DGARL uses a total of 36.21 M parameters, including 3.24 M for the Generator, 7.07 M × 2 for the Discriminator, 6.31 M for the Policy Net, 6.26 M for the Value Net, and 6.26 M for the Soft Q Net. In contrast, during the prediction phase, only 9.55 M parameters are required (the sum of the Policy Net and Generator parameters), which is better than Mask RCNN (43.92 M) and BMA (12.75 M, only Generator). It should be noted that Mask RCNN has a significantly larger number of parameters than other methods, which makes its fitting ability more prominent, resulting in comparable performance to many semi-supervised methods when using only labeled data.

G. Limitations and future direction

As a pioneering work combining DRL with GAN for medical image segmentation, our DGARL still has some limitations: 1) Our method requires the use of reinforcement learning to locate the position of the target through multiple iterations, which is computationally intensive. To address this limitation, we plan to investigate model compression and knowledge distillation techniques to optimize the reinforcement learning model and reduce the computational burden while maintaining the segmentation accuracy. 2) Our method currently starts each training with a brand new GAN, and the time-consuming GAN have led to a less efficient training process of our method. To mitigate this constraint, we are contemplating the integration of pre-trained GAN architectures. By leveraging transfer learning, we can potentially reduce the training time while maintaining, or even enhancing, the reward signal's quality.

V. CONCLUSION

In this study, we proposed a novel semi-supervised method that integrates generative adversarial networks and deep reinforcement learning to achieve accurate segmentation. The proposed method, called DGARL, establishes a reciprocal pipeline that seamlessly integrates DRL with its detection task and GAN with its segmentation task, mutually guiding information between the two tasks to promote each other. Furthermore, our proposed bidirectional exploration DRL alleviates the sparse reward problem of single forward exploration in DRL, while our proposed Stage-confluent GAN provides stable rewards that enhance both detection and segmentation accuracy. The comparison results of the proposed method on two different datasets consistently demonstrate its ability to achieve high-precision segmentation. Overall, our study provides a significant contribution towards developing more efficient and accurate medical image segmentation frameworks.

REFERENCES

[1] D. Zhang, B. Chen, J. Chong, and S. Li, "Weakly-supervised teacher-student network for liver tumor segmentation from non-enhanced images," *Medical Image Analysis*, vol. 70, p. 102005, 2021.

[2] S. K. Zhou, H. N. Le, K. Luu, H. V. Nguyen, and N. Ayache, "Deep reinforcement learning in medical imaging: A literature review," *Medical image analysis*, vol. 73, p. 102193, 2021.

[3] H. Yang, C. Shan, A. F. Kolen *et al.*, "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*. Springer, 2020, pp. 646–655.

[4] Z. Tian, X. Si, Y. Zheng, Z. Chen, and X. Li, "Multi-step medical image segmentation based on reinforcement learning," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–12, 2020.

[5] L. Yu, S. Wang, X. Li, C.-W. Fu, and P.-A. Heng, "Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 605–613.

[6] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[7] A. Casanova, P. O. Pinheiro, N. Rostamzadeh, and C. J. Pal, "Reinforced active learning for image segmentation," *arXiv preprint arXiv:2002.06583*, 2020.

[8] C. Xu, Y. Wang, D. Zhang, L. Han, Y. Zhang, J. Chen, and S. Li, "Bmanet: Boundary mining with adversarial learning for semi-supervised 2d myocardial infarction segmentation," *IEEE Journal of Biomedical and Health Informatics*, 2022.

[9] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," in *International conference on machine learning*. PMLR, 2014, pp. 387–395.

[10] S. Fujimoto, H. Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *International conference on machine learning*. PMLR, 2018, pp. 1587–1596.

[11] R. Jiao, Y. Zhang, L. Ding, R. Cai, and J. Zhang, "Learning with limited annotations: a survey on deep semi-supervised learning for medical image segmentation," *arXiv preprint arXiv:2207.14191*, 2022.

[12] S. Arora and P. Doshi, "A survey of inverse reinforcement learning: Challenges, methods and progress," *Artificial Intelligence*, vol. 297, p. 103500, 2021.

[13] C. Yu, G. Ren, and J. Liu, "Deep inverse reinforcement learning for sepsis treatment," in *2019 IEEE international conference on healthcare informatics (ICHI)*. IEEE, 2019, pp. 1–3.

[14] A. Nair, B. McGrew, M. Andrychowicz, W. Zaremba, and P. Abbeel, "Overcoming exploration in reinforcement learning with demonstrations," in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 6292–6299.

[15] C. Florensa, D. Held, M. Wulfmeier, M. Zhang, and P. Abbeel, "Reverse curriculum generation for reinforcement learning," in *Conference on robot learning*. PMLR, 2017, pp. 482–495.

[16] A. Jabbar, X. Li, and B. Omar, "A survey on generative adversarial networks: Variants, applications, and training," *ACM Computing Surveys (CSUR)*, vol. 54, no. 8, pp. 1–49, 2021.

[17] L. Yu, W. Zhang, J. Wang, and Y. Yu, "Seqgan: Sequence generative adversarial nets with policy gradient," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 31, 2017.

[18] M. Riedmiller, R. Hafner, T. Lampe, M. Neunert, J. Degrave, T. Wiele, V. Mnih, N. Heess, and J. T. Springenberg, "Learning by playing solving sparse reward tasks from scratch," in *International conference on machine learning*. PMLR, 2018, pp. 4344–4353.

[19] N. Souly, C. Spampinato, and M. Shah, "Semi supervised semantic segmentation using generative adversarial network," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5688–5696.

[20] D. C. Elton, Z. Boukouvalas, M. D. Fuge, and P. W. Chung, "Deep learning for molecular design—a review of the state of the art," *Molecular Systems Design & Engineering*, vol. 4, no. 4, pp. 828–849, 2019.

[21] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *International conference on machine learning*. PMLR, 2018, pp. 1861–1870.

[22] S. S. M. Salehi, D. Erdogmus, and A. Gholipour, "Tversky loss function for image segmentation using 3d fully convolutional deep networks," in *Machine Learning in Medical Imaging: 8th International Workshop, MLMI 2017, Held in Conjunction with MICCAI 2017, Quebec City, QC, Canada, September 10, 2017, Proceedings 8*. Springer, 2017, pp. 379–387.

[23] J. Wu, Z. Wei, W. Li, Y. Wang, Y. Li, and D. U. Sauer, "Battery thermal-and health-constrained energy management for hybrid electric bus based on soft actor-critic drl algorithm," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 6, pp. 3751–3761, 2020.

[24] A. Trott, S. Zheng, C. Xiong, and R. Socher, "Keeping your distance: Solving sparse reward tasks using self-balancing shaped rewards," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[25] J. Cheng, W. Huang, S. Cao, R. Yang, W. Yang, Z. Yun, Z. Wang, and Q. Feng, "Enhanced performance of brain tumor classification via tumor region augmentation and partition," *PloS one*, vol. 10, no. 10, p. e0140381, 2015.

[26] Z. Xu, Q. Huang, J. Park, M. Chen, D. Xu, D. Yang, D. Liu, and S. K. Zhou, "Supervised action classifier: Approaching landmark detection as image partitioning," in *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20*. Springer, 2017, pp. 338–346.

[27] Y. Li, N. He, S. Peng, K. Ma, and Y. Zheng, "Deep reinforcement exemplar learning for annotation refinement," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24*. Springer, 2021, pp. 487–496.

[28] M. Tokic and G. Palm, "Value-difference based exploration: adaptive control between epsilon-greedy and softmax," in *KI 2011: Advances in Artificial Intelligence: 34th Annual German Conference on AI, Berlin, Germany, October 4-7, 2011. Proceedings 34*. Springer, 2011, pp. 335–346.

[29] P. Auer, "Using confidence bounds for exploitation-exploration tradeoffs," *Journal of Machine Learning Research*, vol. 3, no. Nov, pp. 397–422, 2002.

[30] H. R. Roth, A. Farag, E. B. Turkbey, L. Lu, J. Liu, and R. M. Summers., "Nih pancreas-ct dataset." [Online]. Available: http://doi.org/10.7937/K9/TCIA.2016.tNB1kqBU

[31] Y. Wu, Z. Wu, Q. Wu, Z. Ge, and J. Cai, "Exploring smoothness and class-separation for semi-supervised medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, vol. 13435. Springer, Cham, 2022, pp. 34–43.

[32] Y. Bai, D. Chen, Q. Li, W. Shen, and Y. Wang, "Bidirectional copy-paste for semi-supervised medical image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 11 514–11 524.

[33] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.
color

[34] D. Müller, I. Soto-Rey, and F. Kramer, "Towards a guideline for evaluation metrics in medical image segmentation," *BMC Research Notes*, vol. 15, no. 1, pp. 1–8, 2022.

[35] O.-M. Pedersen, E. Misimi, and F. Chaumette, "Grasping unknown objects by coupling deep reinforcement learning, generative adversarial networks, and visual servoing," in *2020 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2020, pp. 5655–5662.

[36] Z. Wang, H. Zhu, M. He, Y. Zhou, X. Luo, and N. Zhang, "Gan and multi-agent drl based decentralized traffic light signal control," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 2, pp. 1333–1348, 2021.

[37] J. Ho and S. Ermon, "Generative adversarial imitation learning," *Advances in neural information processing systems*, vol. 29, 2016.

[38] J. Fu, K. Luo, and S. Levine, "Learning robust rewards with adversarial inverse reinforcement learning," *arXiv preprint arXiv:1710.11248*, 2017.

[39] J. Singh and L. Zheng, "Combining semantic guidance and deep reinforcement learning for generating human level paintings," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16 387–16 396.

[40] W. Cui, Y. Liu, Y. Li, M. Guo, Y. Li, X. Li, T. Wang, X. Zeng, and C. Ye, "Semi-supervised brain lesion segmentation with an adapted mean teacher model," in *Information Processing in Medical Imaging: 26th International Conference, IPMI 2019, Hong Kong, China, June 2–7, 2019, Proceedings 26*. Springer, 2019, pp. 554–565.

[41] B. H. Thompson, G. Di Caterina, and J. P. Voisey, "Pseudo-label refinement using superpixels for semi-supervised brain tumour segmentation," in *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2022, pp. 1–5.

[42] N. Souly, C. Spampinato, and M. Shah, "Semi supervised semantic segmentation using generative adversarial network," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5688–5696.

[43] Y. Li, Y. Shen, J. Zhang, S. Song, Z. Li, J. Ke, and D. Shen, "A hierarchical graph v-net with semi-supervised pre-training for histological image based breast cancer classification," *IEEE Transactions on Medical Imaging*, 2023.

[44] H. Basak and Z. Yin, "Pseudo-label guided contrastive learning for semi-supervised medical image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 786–19 797.

[45] J. Cheng, W. Huang, S. Cao, R. Yang, W. Yang, Z. Yun, Z. Wang, and Q. Feng, "Enhanced performance of brain tumor classification via tumor region augmentation and partition," *PloS one*, vol. 10, no. 10, p. e0140381, 2015.

[46] J. Cheng, W. Yang, M. Huang, W. Huang, J. Jiang, Y. Zhou, R. Yang, J. Zhao, Y. Feng, Q. Feng *et al.*, "Retrieval of brain tumors by adaptive spatial pooling and fisher vector representation," *PloS one*, vol. 11, no. 6, p. e0157112, 2016.

[47] C. Xu, Y. Song, D. Zhang, L. K. Bittencourt, S. H. Tirumani, and S. Li, "Spatiotemporal knowledge teacher–student reinforcement learning to detect liver tumors without contrast agents," *Medical Image Analysis*, vol. 90, p. 102980, 2023.

[48] C. Xu, D. Zhang, Y. Song, L. Kayat Bittencourt, S. H. Tirumani, and S. Li, "Contrast-free liver tumor detection using ternary knowledge transferred teacher-student deep reinforcement learning," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 266–275.

[49] A. Hamamci, N. Kucuk, K. Karaman, K. Engin, and G. Unal, "Tumorcut: segmentation of brain tumors on contrast enhanced mr images for radiosurgery applications," *IEEE transactions on medical imaging*, vol. 31, no. 3, pp. 790–804, 2011.

[50] I. Diaz, P. Boulanger, R. Greiner, and A. Murtha, "A critical review of the effects of de-noising algorithms on mri brain tumor segmentation," in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2011, pp. 3934–3937.

[51] H. Greenspan, A. Ruf, and J. Goldberger, "Constrained gaussian mixture model framework for automatic segmentation of mr brain images," *IEEE transactions on medical imaging*, vol. 25, no. 9, pp. 1233–1245, 2006.

[52] S. Zhou, D. Nie, E. Adeli, J. Yin, J. Lian, and D. Shen, "High-resolution encoder–decoder networks for low-contrast medical image segmentation," *IEEE Transactions on Image Processing*, vol. 29, pp. 461–475, 2019.

[53] F. Lyu, M. Ye, A. J. Ma, T. C.-F. Yip, G. L.-H. Wong, and P. C. Yuen, "Learning from synthetic ct images via test-time training for liver tumor segmentation," *IEEE transactions on medical imaging*, vol. 41, no. 9, pp. 2510–2520, 2022.

[54] K. Tong and Y. Wu, "Deep learning-based detection from the perspective of small or tiny objects: A survey," *Image and Vision Computing*, vol. 123, p. 104471, 2022.

[55] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.

[56] Y. Man, Y. Huang, J. Feng, X. Li, and F. Wu, "Deep q learning driven ct pancreas segmentation with geometry-aware u-net," *IEEE transactions on medical imaging*, vol. 38, no. 8, pp. 1971–1980, 2019.

[57] X. Chen, X. Lin, Q. Shen, and X. Qian, "Combined spiral transformation and model-driven multi-modal deep learning scheme for automatic prediction of tp53 mutation in pancreatic cancer," *IEEE Transactions on Medical Imaging*, vol. 40, no. 2, pp. 735–747, 2020.

[58] L. Xie, Q. Yu, Y. Zhou, Y. Wang, E. K. Fishman, and A. L. Yuille, "Recurrent saliency transformation network for tiny target segmentation in abdominal ct scans," *IEEE transactions on medical imaging*, vol. 39, no. 2, pp. 514–525, 2019.