MRL-Seg: Overcoming Imbalance in Medical Image Segmentation with Multi-Step Reinforcement Learning

Feiyang Yang, Xiongfei Li, Haoran Duan, Feilong Xu, Yawen Huang, Xiaoli Zhang, Yang Long, Yefeng Zheng *Fellow, IEEE*

***Abstract*— Medical image segmentation is a critical task for clinical diagnosis and research. However, dealing with highly imbalanced data remains a significant challenge in this domain, where the region of interest (ROI) may exhibit substantial variations across different slices. This presents a significant hurdle to medical image segmentation, as conventional segmentation methods may either overlook the minority class or overly emphasize the majority class, ultimately leading to a decrease in the overall generalization ability of the segmentation results. To overcome this, we propose a novel approach based on multi-step reinforcement learning, which integrates prior knowledge of medical images and pixel-wise segmentation difficulty into the reward function. Our method treats each pixel as an individual agent, utilizing diverse actions to evaluate its relevance for segmentation. To validate the effectiveness of our approach, we conduct experiments on four imbalanced medical datasets, and the results show that our approach surpasses other state-of-the-art methods in highly imbalanced scenarios. These findings hold substantial implications for clinical diagnosis and research.**

***Index Terms*— imbalanced medical image segmentation, deep learning, radiomics, reinforcement learning**

I. Introduction

MEDICAL image segmentation plays a critical role in numerous medical image processing applications, such as structural and functional analysis, diagnosis, and therapy [1], [2]. Every year, kidney tumors affect over 400,000 people worldwide, posing a severe threat to human health [3]. With the advent of computer vision, deep learning-based automatic or interactive segmentation methods have emerged, achieving remarkable success in 2D medical image segmentation [4]. However, in medical images, the regions of interest (ROI) and non-interest regions often exhibit significant imbalances, which are commonly referred to as pixel-level class imbalances. For instance, CT slices of the same patient may display lesions of varying sizes, with some consisting of thousands of pixels while others comprise only a few dozen [5]. Unfortunately, while existing medical image segmentation methods have made significant progress, they have yet to provide a fundamental solution to highly imbalanced segmentations.

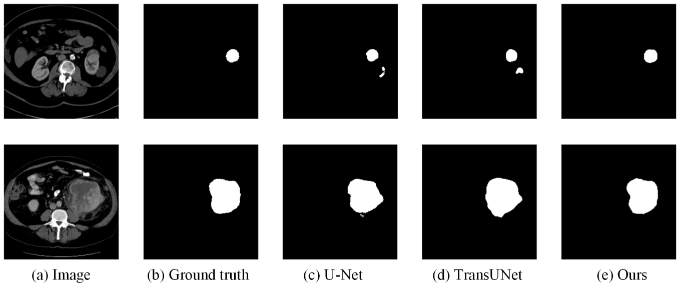
(a) Image (b) Ground truth (c) U-Net (d) TransUNet (e) Ours

Fig. 1. Kidney tumor segmentation results. (a) Raw CT images, where the tumor size varies dramatically across patients/slices and the ratio of non-interest tissue to the tumor is imbalanced with orders of magnitude. (b) Corresponding segmentation labels. (c)-(d) Segmentation results obtained by a U-Net-based method and a Transformer-based method, respectively. (e) Obtained by our MRL-Seg model.

Recent studies in medical image segmentation have proposed various methods to achieve fine results for imbalanced segmentation, with re-weighting and re-sampling being the most commonly used techniques. Along the line of the two aspects, several ideas were proposed involving the network structure [6] [7] [8], the loss function design [9], interactive segmentation [10], and multi-scale segmentation [11] [12] [13]. *Cao et al.* [14] and *Zhang et al.* [15] combined global self-attention mechanisms and Convolutional Neural Networks (CNNs), which relieve the limitations in localization abilities and explicitly modeling long-range dependency. Nevertheless, this strategy merely marginally improves split stability for imbalanced datasets, which focuses more on feature extraction. *Xie et al.* [16] proposed to use a recurrent saliency transformation network to place more emphasis on tiny target regions during adaptation, via complex adversarial training.

Corresponding author: Xiaoli Zhang, Yawen Huang, Haoran Duan (E-mail: zhangxiaoli@jlu.edu.cn; yawenhuang@tencent.com; haoran.duan@ieee.org).

Feiyang Yang, Xiongfei Li, Feilong Xu and Xiaoli Zhang are with Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University; the College of Computer Science and Technology, Jilin University, Changchun, China (E-mail: yangfy21@mails.jlu.edu.cn; lxf@jlu.edu.cn; xufl21@mails.jlu.edu.cn; zhangxiaoli@jlu.edu.cn).

Yawen Huang and Yefeng Zheng are with Tencent Jarvis Lab, Shenzhen, China (E-mail: yawenhuang@tencent.com; yefengzheng@tencent.com).

Haoran Duan and Yang Long are with the Department of Computer Science, Durham University (E-mail: haoran.duan@ieee.org; yang.long@durham.ac.uk).

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2023.3336726

This method has achieved good segmentation for small targets, yet failed for scenarios where small targets and large targets need to be segmented at the same time. *Zhang et al.* [17] proposed a new weighting strategy for weighting the attention matrix in the Transformer for improving the accuracy and robustness of newborn brain MRI image segmentation. *Di et al.* [18] proposed a hybrid end-to-end TD-Net for automatic liver tumor segmentation from CT images which utilizes a novel multi-scale contextual attention mechanism. *Man et al.* [19] proposed to use geometry awareness to locate the lesion area by deep Q-learning, which obtains more accurate localization of the lesion area. This two-stage model of first localization and then segmentation is strenuous to learn directly to adapt to the size of the segmentation window.

Despite the success achieved by existing approaches, these methods may not provide sufficient robustness to challenging regions in 2D medical image segmentation. Deep learning-based medical image segmentation methods typically rely on minimizing cross-entropy (CE) during training. The CE loss function calculates the similarity between the probability distributions of the predicted segmentation and the corresponding ground truth. Employing the CE loss function assumes equal contributions from all samples and classes to the training loss, thereby necessitating a comprehensive dataset and well-balanced classes to achieve robust generalization. However, the efficacy of CE can be limited in a highly imbalanced setting. Specifically, imbalanced medical image datasets exhibit two main characteristics: (1) Inter-class-imbalance during training, i.e., much fewer foreground pixels relative to a large number of background pixels in binary segmentation; (2) Significant discrepancy between foreground objectives, i.e., differences in the number of foreground pixels between segmentation targets. The classes with more observations (i.e., pixels) overshadow the minority classes.

Although cropping or localization can be used to constrain the foreground/background ratio, there are still two issues: (1) When a uniform scale is applied for cropping, we are faced with the task of managing foregrounds (lesions) of diverse scales; (2) When various scales are utilized for cropping, we must handle segmentation windows of differing sizes. In practice, it proves challenging to first categorize lesion sizes and then apply corresponding segmentation. As shown in Fig. 1(a), the two raw images are kidney CT from different slices of the same patient, and the size of the region of interest in the images varies significantly. Performing segmentation at different scales via the same size segmentation window is of great challenge. As illustrated in Fig. 1(c) and (d), U-Net-based models always incur positioning offset, causing the lesion outline to deviate from the ground truth; Transformer-based models exhibit low stability of edge segmentation. In addition, it should be noted that both methods have a tendency to mistakenly identify non-lesional tissue in neighboring organs as lesions, which is a type of error that is less commonly made by human annotators.

To tackle the difficulties in highly imbalanced medical image segmentation, we propose a multi-step reinforcement learning for medical image segmentation, called MRL-Seg. The advantage of our method is to achieve excellent performance on imbalanced data without modifying complex loss functions or training adaptively sized segmentation windows. As shown in Fig. 1(e), our model outperforms previous work with targeted adjustments for this type of sample. We model a patch-based representation learning for medical images as a Markov decision process (MDP) and solve it by employing multi-step reinforcement learning (MRL). To shrink the exploration space to an acceptable size, each patch needs to be treated as an agent with a shared pixel-level behavior policy. We set the segmentation result as the environmental reward and set the action as retain or mask for each patch. As the update of the network, more pixels that are not conducive to segmentation are masked, and the imbalanced categories gradually tend to be balanced, which reduces the distractors for imbalanced image segmentation. In order to make the results obtained by the segmentation model closer to the judgment of human experts, we use two kinds of prior knowledge as the theoretical basis of MRL reward function design, consisting of measures of the ROI scale and the approximate difficulty of segmentation. Our results on KiTS19 and JLUKT datasets demonstrate that the proposed method significantly outperforms the existing methods, achieving more satisfactory segmentation performances.

To summarize, our contributions are shown as follows:

- We propose a novel approach, MRL-Seg, which integrates three structural-oriented modules into a unified network to achieve robust adaptation for highly imbalanced medical image segmentation.
- We introduce a pioneering approach for patch representation based on multi-step reinforcement learning, which can intelligently and effectively mitigate the severe effects of differences spanning several orders of magnitude between regions of interest and non-interest regions on lesion segmentation.
- We incorporate two types of prior knowledge into the reward function design of our reinforcement learning approach, which remarkably enhances the localization accuracy of lesion outlines and ensures the stability of boundaries.
- We conduct comprehensive experiments on four imbalanced medical datasets. Experimental results demonstrate the significant superiority of our MRL-Seg to multiple state-of-the-art approaches.

II. Related Work

A. Automatic 2D Medical Image Segmentation with Deep Learning

For automatic segmentation, existing deep learning approaches can be grouped into two categories, i.e., U-Net-based and Transformer-based: 1) A U-Net-based method exploits a symmetrical U-shaped structure. By encoding and decoding image features, the network fuses both high-level and low-level features in the network to obtain better segmentation effects. 2) A Transformer based method utilizes self-attention to replace classic Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs), adopting the strategy of stacking self-attention and fully connected layers that possess a powerful global information induction ability.

U-Net-based Segmentation. One common limitation of CNN-based models is their difficulty in capturing long-range dependencies or relations between input features, particularly for target structures that display significant inter-patient variability in terms of texture, shape, and size. To tackle this, the U-shaped design has emerged as the go-to approach for numerous image segmentation tasks, owing to its remarkable success [1]. The conventional U-Net introduces an encoder-decoder structure, where the encoder is used to extract features from the input image, and the decoder is used to transform these features into a segmentation result. The architecture also includes skip connections that connect corresponding features from the encoder to the decoder, preserving low-level feature information. Consequently, a plethora of derived algorithms, such as Dense U-Net [20], ResU-Net [21], and nnU-Net [4], have been proposed. U-Net and its variants have become a benchmark model for various image segmentation tasks which can be trained end-to-end through fine-tuning without the need for manually designing feature extractors, making it easier to use and implement.

Transformer-based Segmentation. Transformer [22] is commonly adopted in natural language processing. Different from the previous encoder-decoder frameworks, Transformer utilizes self-attention to replace classic CNNs or RNNs, adopting the strategy of stacking self-attention and fully connected layers. Driven by the success of Transformer, *Dosovitskiy et al.* [23] first introduced a pioneering Vision Transformer (ViT) with excellent performance in image recognition tasks. Thereafter, to apply Transformer to the application of medical image segmentation, several modifications have been proposed. For instance, [24] is the first model designed for the medical image segmentation, establishing a self-attention mechanism from a sequence-to-sequence prediction perspective. A hybrid CNN-Transformer architecture is adopted by TransUNet to exploit detailed high-level spatial information from CNN features and the global context encoded by Transformer. Nonetheless, the samples of the same disease vary greatly among different patients (e.g., regions of interest of various sizes), which limits the application of the Transformer in medical image segmentation.

To tackle the paucity of long-range dependencies of U-Net-based methods and the lack of inductive bias of Transformer-based methods, several works were proposed, inspired by the U-shaped design and Vision Transformer, For example, *Xie et al.* [25] proposed a novel medical image segmentation called DMCGNet, which combines dense self-mimic and channel grouping mechanisms to solve the class imbalance of medical image segmentation. *Li et al.* [26] proposed a hierarchical U-Net for the segmentation of ovaries and follicles in ultrasound images. *Cao et al.* [14] used a hierarchical Swin Transformer with an offset window as the encoder to extract contextual features, and designed a symmetric Swin Transformer-based decoder with patch expansion layers by upsampling operations to recover the spatial resolution of feature maps. *Yuan et al.* [12] used the spatial attention mechanism and the channel attention mechanism to enhance the capture of blood vessel features in medical images. *Zhang et al.* [15] used Transformer and CNNs in parallel to produce great segmentation results for polyp, skin lesion, hip, and prostate on both 2D and 3D medical images. *Valanarasu et al.* [27] realized the improvement in performance without any need for pre-training by the design of a gated axial attention layer to explore the feasibility of applying the Transformer in the light of only self-attention.

B. Vision in Reinforcement Learning

Reinforcement learning (RL) studies the problem of how to maximize rewards in a complex and uncertain environment. RL involves the model taking actions in an environment and receiving rewards or penalties based on the success of those actions. By learning from these rewards and penalties, the model can adjust its behavior and decision-making processes. Existing RL approaches mainly focus on policy-based agents, value-based agents, and actor-critic agents. We have collectively witnessed some breakthroughs in the fields of game playing, robotics, autonomous driving, computer vision, and natural language processing in recent years. The most influential breakthroughs include DQN [28] and DeepStack [29], pioneering a new era of deep reinforcement learning. Indeed, RL is an efficient tool to address sequential decision-making problems.

However, in the domain of computer vision, the application of RL is confronted with some practical challenges, such as image cropping [30] and global color enhancement [31]. RL cannot effectively deal with tasks requiring pixel-level operations, such as image representation learning. In order to achieve even more optimization, *Furuta et al.* [32] originally suggested using reinforcement learning for pixel-level image processing tasks such as local color improvement, picture restoration, and image denoising. The agents learn the best course of action for increasing the mean expected total return across all pixels because the number of agents and pixels is equal. Each pixel's value is taken to represent the current situation and is then changed by agents' iterative activities. Therefore, the actions performed by the agent are understandable to humans, which is tremendously different from traditional CNNs. Based on this, *liao et al.* [33] proposed a multi-agent reinforcement learning (MARL)-based 3D medical image voxel interactive segmentation framework (IteR-MRL). The model adapts the concept of reinforcement learning to user interaction, utilizing a pre-detection of lesions as an input to the algorithm. Apart from this, *Man et al.* [19] and *Tao et al.* [34] attempted to utilize reinforcement learning to initially identify the approximate location of the ROIs and subsequently perform precise segmentation. This two-stage model heavily relies on the accuracy of the initial positioning, and achieving adequate generalization can be challenging when dealing with complex ROI shapes.

III. Method

A. Overview

The overall process is illustrated in Fig. 2, comprising three modules: Reinforcement-Net, Representation-Net, and

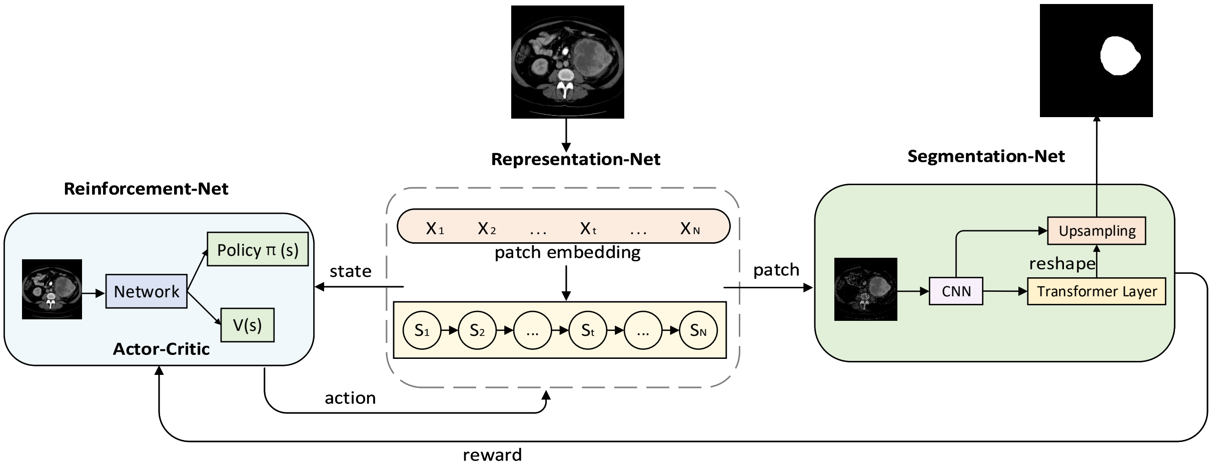

Fig. 2. Illustration of the overall process. At each state, the Reinforcement-Net samples an action, while the Representation-Net provides the state to the Reinforcement-Net and produces the final image representation to the Segmentation-Net once all actions are sampled. The Segmentation-Net then generates results and supplies reward values to the Reinforcement-Net.

Segmentation-Net. Reinforcement-Net is an RL-Based Asynchronous Advantage Actor-Critic (A3C) [35] module that transforms pixel-level segmentation into a sequential decision task and enriches the prior knowledge for highly imbalanced segmentation by interleaving updating of policy and values. Representation-Net is responsible for updating the images based on reinforcement learning and serves as input to Segmentation-Net. Segmentation-Net, a U-shaped module that combines CNNs and Transformer, produces outputs that are fed into Reinforcement-Net, where the reward is ultimately calculated. By continuously updating Representation-Net, Segmentation-Net achieves better predictions slice by slice, while Reinforcement-Net is constantly refined with improved policies. These three modules operate in a closed loop to facilitate seamless integration.

Specifically, we denote Policy $\pi(S)$ as the update strategy and $V(S)$ as the value of the reward function. We adopt the A3C as the framework for the Reinforcement-Net, taking each pixel as the state of the agent, and the effect of segmentation as the reward of the environment to guide each agent to actions that are beneficial to the segmentation task. For the state design, we define I_i as the i-th pixel of the input image I, with a total of N pixels ($i = 1, \cdots, N$). Each pixel is regarded as a single agent and the updated policy is represented as $\pi_i\left(a_i^{(t)} \mid s_i^{(t)}\right)$. For the action design, $a_i^{(t)} \in \mathcal{A}$ represents the action of the i-th agent at step t, $\mathcal{A}$ is the action space $\{mask, retain\}$. The actions can be denoted as $\boldsymbol{a}^{(t)} = \left(a_1^{(t)}, \cdots, a_N^{(t)}\right)$. For the whole image, the state from the Representation-Net can be represented as $\boldsymbol{s}^{(t+1)} = \left(s_1^{(t+1)}, \cdots, s_N^{(t+1)}\right)$. The reward from the Segmentation-Net can be denoted as $\boldsymbol{r}^{(t)} = \left(r_1^{(t)}, \cdots, r_N^{(t)}\right)$. We use the policy of iterative learning $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_N)$ to maximize the mathematical expectation of the reward value of all pixels. Since the size of the final fully connected layer of a single agent is $|\mathcal{A}|^N$, it is hard to calculate since N is a huge number. Therefore, we follow the method of PixelRL [32], where N agents can share parameters by using FCNs, and our method can handle images with different sizes.

To adapt the RL training, the Reinforcement-Net first uses three convolution blocks and one Transformer block to extract high-level features as shown in Fig. 3. The network includes two kinds of heads: a policy head and a value head, both of which are composed of three convolution compositions. The convolution blocks and Transformer block in the feature extraction part have the same structure as the Segmentation-Net. Furthermore, the Segmentation-Net is pre-trained to ensure the stability and convergence of the Reinforcement-Net during its training. Specifically, the feature extraction blocks of the Reinforcement-Net are initialized using the weights of the corresponding blocks from the pre-trained Segmentation-Net. The numbers of output channels of the last convolutional layer in the Actor-Net are the same as the number of actions.

Indeed, medical images tend to have stronger prior knowledge than natural images, e.g., we know the approximate location of the lesion or organ as well as the size of the lesion before the model training. From these two aspects, we design a novel reinforcement learning reward function based on prior knowledge, which is detailed in Sections III-B and III-C, respectively. In Section III-D, we present our Segmentation-Net.

B. Optimize imbalanced Segmentation via Learning the Foreground-Background Ratio

As mentioned in Section I, the ratio of the region of interest to other parts is crucial for the segmentation result. For instance, in an image of a kidney tumor from a single patient, there may exist a substantial difference between the largest and smallest lesion areas, resulting in a significant imbalance (which varies across slices) between the foreground and background of the segmentation. In our case, given the original image, we set the actions set as $\{mask, retain\}$, i.e., masking the parts that have adverse effects on the segmentation and retaining the parts that are beneficial to the segmentation.

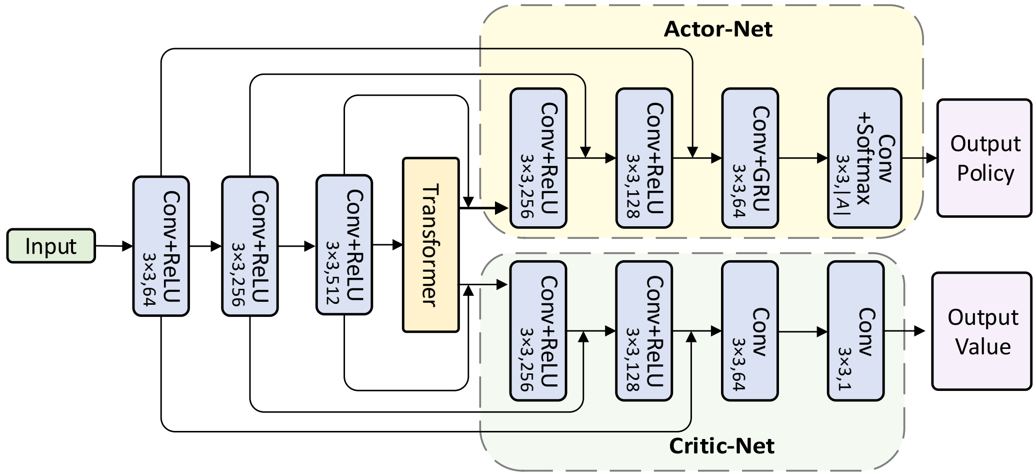

Fig. 3. The network architecture for Reinforcement-Net. The policy and value heads share the low-level features and extract their own high-level features.

Meanwhile, we set the reward as the calculated score from the Segmentation-Net.

As shown in Fig. 3, we apply Actor-Critic, which comprises a policy-based Actor-Net and a value-based Critic-Net. The Actor-Net is responsible for predicting the actions made based on the patch, and the Critic-Net evaluates how much the segmentation score has been improved given the current action sequence. Both networks employ the state $s^{(t)}$ as the input of the model, where $s^{(t)}$ is the state at step t, and the Critic network outputs the $V\left(s^{(t)}\right)$, which is the expected total reward of $s^{(t)}$. The Actor-Net produces policy as $\pi\left(a^{(t)} \mid s^{(t)}\right)$. Our approach aims to discover the optimal policies $\pi = (\pi_1, \cdots, \pi_N)$ that maximize the mean of the total expected rewards at all pixels:

$$\boldsymbol{\pi}^* = \underset{\boldsymbol{\pi}}{\operatorname{argmax}} \mathbb{E}_{\boldsymbol{\pi}} \left(\sum_{t=0}^{\infty} \gamma^t r^{(t)} \right), \tag{1}$$

$$\bar{r}^{(t)} = \frac{1}{N} \sum_{i=1}^{N} r_i^{(t)}, \tag{2}$$

where $r_i^{(t)}$ denotes the reward at each pixel, and $\bar{r}^{(t)}$ denotes the mean of rewards. We set each pixel (represented by a patch centered on it) as an agent and construct a fully convolutional A3C network in the form of FCNs.

In practice, kidney tumor image segmentation can be regarded as a binary classification of pixels. Due to the inhomogeneity of the categories, the pixel classification is unstable. Hence, we embed segmentation prior knowledge into the global reward, by rewriting Eq. (2) as follows:

$$\bar{r^{*}}^{(t)} = \frac{1}{N} \left(\sum_{i=1}^{N} r_i^{(t)} + \alpha \left(\frac{N_{\text{Front}}}{N_{\text{Back}}} + \beta \cdot \frac{N_{\text{Back}}}{N_{\text{Front}}} \right) \right), \tag{3}$$

where $\alpha \in (-1, 0)$ and $\beta \in (0.2, 0.3)$ are hyperparameters balancing the weights of different terms. In binary classification tasks (Regions of Interest and Non-Regions of Interest), the optimal balance between positive and negative samples can depend on the specific context and the model being used. However, a common guideline is to aim for a roughly equal distribution of positive and negative samples [36]. Let $x = \frac{N_{\text{Front}}}{N_{\text{Back}}}$, we use the function $f(x) = x + \beta/x$, which is a unimodal function. Our goal is for x to reside within the range of 0.45 to 0.55 when $f(x)$ reaches its extremal values. To achieve this, we adjust the value of β to lie between 0.2 and 0.3. Thus, the ratio of foreground to background is constrained to be between 44.7% and 54.8%. With the constraints of this prior knowledge, patches which unconducive to segmentation are masked, and the distribution of foreground and background gradually changes from imbalanced to balanced.

C. Reward Design Based on Segmentation Approximate Difficulty of the Pixels (ADP)

Our local reward design logic for a single agent is based on cross-entropy (CE). In particular, the reward $\mathcal{P}_i$ is designed as the difference between the segmentation prediction and ground truth. To guide the model training in a constrained direction, we refer to the score of the previous step as a comparison and transcendence. For the benchmark, the oscillation in CE value determines whether the agent gets a positive or a negative reward. The reward r_i of each agent is defined as follows:

$$\mathcal{P}_i^{(t)} = -y_i \log\left(p_i^{(t)}\right) - (1 - y_i) \log\left(1 - p_i^{(t)}\right), \tag{4}$$

$$r_i^{(t)} = \mathcal{P}_i^{(t-1)} - \mathcal{P}_i^{(t)}, \tag{5}$$

$$R_i^{(t)} = r_i^{(t)} + \gamma V\left(s_i^{(t+1)}\right), \tag{6}$$

$$A\left(a_i^{(t)}, s_i^{(t)}\right) = R_i^{(t)} - V\left(s_i^{(t)}\right), \tag{7}$$

where y_i denotes the label of the i-th pixel, γ denotes a discount factor, $R_i^{(t)}$ denotes the local reward, and $A\left(a_i^{(t)}, s_i^{(t)}\right)$ denotes the advantage function in A3C for one-step learning. The gradient for each network parameter is the average of the gradients of all pixels as follows:

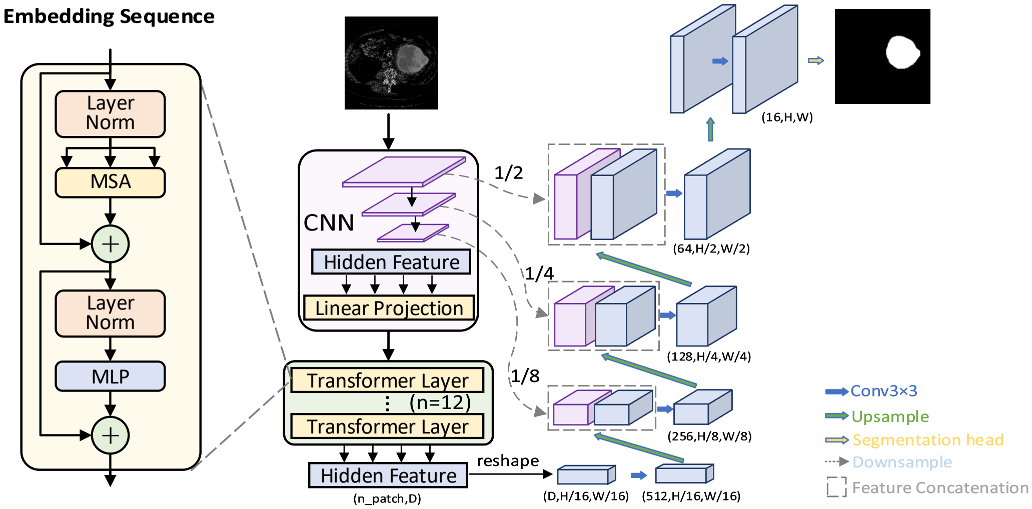

Fig. 4. Segmentation-Net employs a hybrid CNN-Transformer architecture as the encoder as well as a cascaded upsampler to enable precise localization.

$$d\theta_A = -\nabla_{\theta_A}\frac{1}{N}\sum_{i=1}^{N}\log\pi\left(a_i^{(t)} \mid s_i^{(t)}\right)\left(R_i^{(t)} - V\left(s_i^{(t)}\right)\right), \tag{8}$$

$$d\theta_C = \nabla_{\theta_C}\frac{1}{N}\sum_{i=1}\left(R_i^{(t)} - V\left(s_i^{(t)}\right)\right)^2, \tag{9}$$

where θ_A and θ_C denote the policy head and the value head, respectively, from Reinforcement-Net. The gradient of the parameters of the Actor-Net and Critic-Net is the average of the gradients of all patches.

As the receptive field increases, the Actor-Net and the Critic-Net can not only observe the $s_i^{(t)}$ of a single pixel but also observe the adjacent pixels. Moreover, when the receptive field of FCNs is $1*1$, N agents are completely independent. The set of neighbors of a pixel can characterize the surrounding structure of this pixel [37]. We set the approximate difficulty measure of the segmentation to explore neighbor pixel values V. Given the collection of partial pixels $D=\{(x_1,y_1),(x_2,y_2),\ldots,(x_N,y_N)\}$, where x_i denotes an input pixel, $y_i\in\{-1,+1\}$ denotes the corresponding label, each x_i $(i=1,2,\ldots,N)$ corresponds to a partial receptive window $F_i=\{(x_p,y_p)\mid p=1,2,\ldots,P\}$. The segmentation approximate difficulty of the pixels (ADP) is represented as:

$$\mathfrak{J}_i = \frac{\sum_{X_p\in F_i}\mathbb{I}\left(y_p \neq y_i\right)}{|F_i|}, \tag{10}$$

where $|F_i|$ denotes the number of pixels in the partial receptive field. Here, $\mathfrak{J}_i\in[0,1]$. Obviously, $\mathfrak{J}_i=0$ if the labels of all pixels in F_i are the same as x_i, and $\mathfrak{J}_i=1$ if the class labels of all pixels in F_i are different from x_i. Therefore, we can rewrite $R_i^{(t)}$ in Eq. (6) as follows:

$$R_i^{(t)} = r_i^{(t)} + \gamma\sum_{j\in N(i)} w\left(\mathfrak{J}_i + V\left(s_j^{(t+1)}\right)\right), \tag{11}$$

where $N(i)$ is the partial receptive window centered on the i-th pixel, and w represents a convolution filter weight, which cooperates with $\mathfrak{J}_i$ to control the impact of neighboring pixel values in the time $t+1$, the second term in Eq. (11) is derived from a 2D convolution. Thus, in the n-step case, $\boldsymbol{R}^{(t)}$ can be defined as:

$$\begin{aligned}\boldsymbol{R}^{(t)} &= \boldsymbol{r}^{*(t)} + \gamma\boldsymbol{w}*\boldsymbol{r}^{*(t+1)} + \gamma^2\boldsymbol{w}^2*\boldsymbol{r}^{*(t+2)} + \cdots \\ &+ \gamma^{n-1}\boldsymbol{w}^{n-1}*\boldsymbol{r}^{*(t+n-1)} + \gamma^n\boldsymbol{w}^n * V\left(s^{(t+n)}\right),\end{aligned} \tag{12}$$

where $*$ represents convolution operation, $\boldsymbol{w}^n*\boldsymbol{r}^*$ represents the n times convolution on r^* with the convolution filter $\boldsymbol{w}$. The gradient of $\boldsymbol{w}$ is calculated as follows:

$$\begin{aligned}d\boldsymbol{w} = &-\nabla_{w}\tfrac{1}{N}\textstyle\sum_{i=1}^{N}\log\pi\left(a_i^{(t)}\mid s_i^{(t)}\right)\left(R_i^{(t)}-V\left(s_i^{(t)}\right)\right) \\ &+\nabla_{w}\tfrac{1}{N}\textstyle\sum_{i=1}^{N}\left(R_i^{(t)}-V\left(s_i^{(t)}\right)\right)^2.\end{aligned} \tag{13}$$

The w in Eq. (11) can be learned simultaneously with the network parameters θ_A and θ_C in Eq. (8) and (9), due to the mutual containment of $\mathfrak{J}_i$ and w. Note, R_i will not deviate from the predicted value of $V\left(s_i^{(t)}\right)$, making the model easier to converge.

D. Segmentation-Net Module

As discussed in Section II-A, U-Net and Transformer structures can complement each other. Single convolution operation can be regarded as special multi-head self-attention (MSA), whose receptive field is partial, and the attention weights are fixed. Conversely, MSA can also be regarded as a special convolution, whose window size is the whole image, and the summation operation is dynamically weighted according to attention weights. Transformer can alleviate the remote dependence limitation of U-Net. Meanwhile, U-Net can alleviate the lack of low-level details of the Transformer. Our segmentation module can combine the advantages of both.

Given an image $\chi\in\mathbb{R}^{H\times W\times C}$, where H, W, and C denote the image height, width, and channel number, respectively, our goal is to predict the pixel-level binary segmentation label with the corresponding size of $H \times W$. We encode the spatial information of a patch to get the feature representation based on reinforcement learning, and the downsampling part of the segmentation model is designed by the encoder of Transformer, and the upsampling part consists of multi-step decoding of hidden features to output the final segmentation mask. As shown in Fig. 4, we also use the skip-connection to form a U-shaped structure, so as to combine the MSA structure based on Transformer with the CNN structure, making them complement each other.

We map patch χ_p to a latent D-dimensional embedding space, encode the patch spatial information, and add position information to patch embedding $\mathbf{z}_0 = [\chi_p^1 \upsilon, \chi_p^2 \upsilon, \cdots, \chi_p^N \upsilon] + \upsilon_{pos}$, where υ denotes the patch embedding projection, and υ_{pos} denotes the position embedding. After the image is trained by the Reinforcement-Net, the output of the representation network is used as the input of the Segmentation-Net. The encoder of the Transformer consists of L-layer MSA and Multilayer Perceptron (MLP).

IV. Experiments

To verify the effectiveness and robustness of the proposed method, we conduct extensive experiments, which include comparison experiments with state-of-the-art algorithms and sufficient ablation experiments. Besides, the datasets, image quality metrics, and experimental settings used for the experiments are also described in detail.

A. Datasets

KiTS19: This dataset was obtained from the KiTS19 challenge [38], containing 210 cases of kidney tumors with accompanying clinical context, CT semantic segmentation, and surgical outcomes. It was compiled from patients who underwent partial or radical nephrectomy for one or more kidney tumors at the University of Minnesota Medical Center between 2010 and 2018. Out of 300 patients with comprehensive clinical results, 210 cases were randomly selected for inclusion in the dataset.

JLUKT: This medical image dataset was contributed by the Second Affiliated Hospital of Jilin University. The data comprises plain and contrast-enhanced CT scans of both left and right kidneys in various phases, including plain phase, corticomedullary phase, nephrographic phase, and excretion phase, which contains 61 cases of kidney tumor CT images. This dataset will be made publicly available once the paper is accepted for publication.

NIH: The NIH Pancreas image dataset comprises 82 contrast-enhanced abdominal CT scan volumes [39]. Each CT scan boasts a resolution of $512 \times 512 \times L$, where $L \in [181, 466]$ is the number of sampling slices taken lengthwise along the body. The thickness of these slices varies, with measurements ranging from 0.5 mm to 1.0 mm.

BUSI: The dataset comprises 780 images, sourced from two different types of ultrasound equipment - the LOGIQ E9 Ultrasound and the LOGIQ E9 Agile Ultrasound System, both utilized at Baheya Hospital [40]. These images exhibit an average size of 500×500. We apply 487 benign and 210 malignant images.

We compute the foreground-background ratios of KiTS19, JLUKT, and BUSI and divided them into six intervals to investigate the effectiveness of our method more precisely. Given that the size of the region of interest in the NIH dataset is notably consistent, and that the imbalance predominantly manifests in the disparity between the foreground and the background dimensions, we utilized a single ratio interval between the foreground and background in subsequent experiments. The primary characteristics of imbalanced medical image datasets are twofold: (1) the presence of inter-class imbalance during the training phase, manifested as a significantly smaller quantity of foreground pixels compared to the abundant background pixels in binary segmentation (e.g., NIH: the smallest foreground measures merely 2px, while its corresponding background stands at a substantial 50174px.); (2) A noticeable disparity between foreground objectives (e.g., KiTS19: ranging from the smallest lesions (6px) to the largest (30025px) ones.), demonstrated by the varying quantities of foreground pixels among different segmentation targets.

B. Implementation Details

To demonstrate the practical value of the model, the aforementioned four datasets are all utilized with their original dimensions as inputs to the model, and testing is conducted without lesion masks or pre-detection. Notably, no organ labels were utilized during any stage of the process, encompassing both the training and the testing phases. The four datasets were initially resized to a standardized dimension of 512×512 pixels, followed by clipping the CT Hounsfield unit values to the range of $[-79, 304]$. The pixel values were then rescaled to $[0, 255]$. To initialize the Reinforcement-Net, we utilized the pre-trained Segmentation-Net parameters. The segmentation network was trained for 100 epochs with a batch size of 24 using the Adam optimizer. The learning rate was set to 0.01 and followed a step-size decay scheme during training. In contrast to the Segmentation-Net, the Actor-Net of the Reinforcement-Net was equipped with a convolutional layer with an action space, and the last layer of the Critic-Net was replaced with a convolutional layer with one output channel. During the training process, a batch size of 10 was utilized, and the number of epochs was set to 50. In each epoch, 4 actions were performed on each slice before ending it. The discount factor was set to $\gamma = 0.95$. We divide the datasets into four fixed folds, each containing approximately the same number of instances. Following the cross-validation methodology, we train the model on three out of the four subsets and validate it against the remaining one.

C. Evaluation Metrics

1) Dice coefficient: The Dice coefficient is a set similarity measurement function, which mainly calculates the overlap ratio between the segmentation result and the ground truth. The range of values is $[0, 1]$. Specifically, it can be expressed as $\text{Dice}(A, B) = \frac{|A \cap B|}{(|A|+|B|)/2}$, where A represents the segmentation result, and B represents the ground truth. $|A|$ and $|B|$ are the number of elements in A and B, respectively.

TABLE I

THE SEGMENTATION RESULTS (MEAN ± STD) FROM STATE-OF-THE-ART METHODS ON THE KITS19 DATASET ARE SHOWN. THE BEST RESULTS ARE IN BOLD. ASTERISKS MARK SIGNIFICANT DIFFERENCES FROM OUR METHOD, AS PER A PAIRED STUDENT'S T-TEST. (*: P<0.05).

Method	Dice (%) at various foreground-background ratios					
	0%∼2%	2%∼4%	4%∼6%	6%∼8%	8%∼10%	Avg
ResU-Net [21]	52.32±5.71	71.09±6.52	72.41±5.77	72.14±3.63	71.09±4.12	59.15±6.20
nnU-Net [4]	59.71±4.25*	81.06±4.46*	**82.71±4.57**	**83.59±4.03**	81.56±3.56*	67.56±4.77*
Swin-Unet [14]	57.61±5.27	80.01±4.76	79.88±5.23	80.19±4.51	81.02±4.10	65.58±5.42
TransUNet [24]	55.29±5.13	77.61±4.79	76.52±4.16	78.41±4.66	78.22±4.03	63.17±5.12
TransFuse-S [15]	54.10±5.71	77.19±3.96	77.42±3.71	76.53±4.29	78.99±4.19	62.32±5.67
RSTN [16]	56.65±4.25	76.37±4.16	78.96±3.57	79.55±4.13	80.17±3.87	64.06±5.51
UFL [41]	55.19±4.75	76.79±4.11	77.69±5.22	78.21±4.71	78.92±3.58	63.03±5.13
Ours	**61.54±4.63**	**82.71±4.23**	81.06±3.88	83.20±3.72	**82.76±3.33**	**68.94±4.19**

Method	HD at various foreground-background ratios					
	0%∼2%	2%∼4%	4%∼6%	6%∼8%	8%∼10%	Avg
ResU-Net [21]	20.31±6.78	17.69±5.22	18.79±6.31	17.27±6.54	16.61±5.93	19.44±7.30
nnU-Net [4]	16.32±6.13*	**15.14±5.22**	**15.10±5.11**	14.93±4.79*	15.10±4.71*	15.88±6.32*
Swin-Unet [14]	17.90±6.54	16.31±5.76	16.32±5.33	15.71±6.01	16.30±5.12	17.30±6.51
TransUNet [24]	18.72±6.17	17.64±5.21	16.53±4.97	17.22±5.13	16.53±5.42	18.20±6.90
TransFuse-S [15]	18.10±5.97	17.96±6.10	16.72±5.13	17.36±6.10	16.32±5.49	17.89±6.96
RSTN [16]	17.95±6.10	16.77±5.96	17.31±6.07	17.23±5.71	16.43±4.71	17.59±6.62
UFL [41]	17.13±5.77	16.93±6.17	17.41±5.93	16.59±4.79	16.19±5.78	17.07 ±6.71
Ours	**15.34±5.23**	15.26±5.61	15.17±4.96	**13.26±4.26**	**13.57±3.67**	**15.20±5.96**

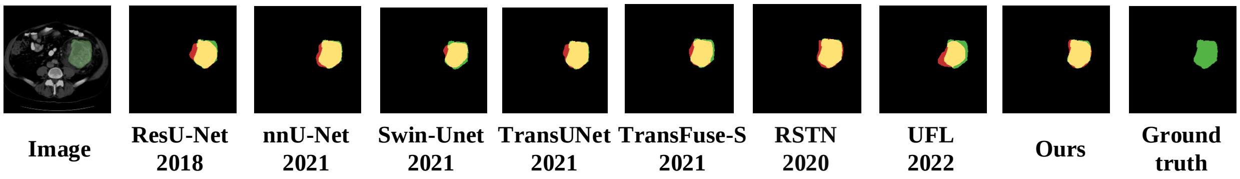

Fig. 5. Visual comparison of lesion segmentation results obtained from various methods on CT images from KiTS19. Red, green, and yellow indicate the prediction, ground truth, and overlapped pixels, respectively.

2) Hausdorff distance: Hausdorff distance evaluates the symmetric distance between the sample X and the sample Y, where x and y are the points of the sets X and Y, respectively, $d(x,y)$ is regarded as the metric between these points, and thus we can take of $d(x,y)$ as the Euclidean distance. It is defined as $\text{Hausdorff} = \max\{d_{XY}, d_{YX}\} = \max\{\max_{x\in X}\min_{y\in Y} d(x,y), \max_{y\in Y}\min_{x\in X} d(x,y)\}$. The Hausdorff distance is more sensitive to the segmented boundary.

D. Comparison with State-of-the-Art

We evaluate the effectiveness of MRL-Seg on two different datasets, namely KiTS19 and JLUKT. Specifically, the proposed method is comprehensively compared with those state-of-the-art methods including: **ResU-Net** [21] uses ResNet block for downsampling and ordinary bi-convolution block for upsampling. **nnU-Net** [4] is a cascade of U-Net, focusing on image preprocessing and the selection of training optimizers. **Swin-Unet** [14] employ a layered Swin Transformer with an offset window as an encoder to extract context features, and it is a pure Transformer similar to U-Net for medical image segmentation. **TransUNet** [24] uses Transformer as the encoder of medical image segmentation task by recovering local spatial information. **TransFuse** [15] propose a novel architecture that ingeniously fuses CNNs and Transformer models to enhance the performance and accuracy of medical image segmentation. **RSTN** [16] proposes a neural network model based on recurrent attention transformation for the segmentation of tiny objects in abdominal CT scans, which is capable of effectively identifying and segmenting small lesions or anomalies. **UFL** [41] generalizes the Dice and the cross-entropy losses, which proposes a unified focal loss to address the problem of class imbalance in medical image segmentation.

To ensure a fair comparison, we only utilized the original images and lesion labels available in the dataset during the experimental evaluation. We used the source code provided by the authors to reimplement [21], [4], [14], [24], [16] under the same experimental configuration as our method, and the corresponding quantitative and qualitative results were given. For the method of [41] and [15] which do not provide source codes, we conducted the experiments according to the methods described in the original paper. Furthermore, we used the same backbone for our method and the reimplemented method of [24]. For all experiments, we adhered to splitting of the training and test data for all comparison methods. Furthermore, we conducted a paired Student's t-test with the second rank results. A p-value less than 0.05 indicates a significant difference between our method and the comparison methods.

The quantitative results on the KiTS19 dataset are presented in Table I. Two observations can be drawn from the results.

TABLE II

THE SEGMENTATION RESULTS (MEAN ± STD) FROM STATE-OF-THE-ART METHODS ON THE JLUKT DATASET ARE SHOWN. THE BEST RESULTS ARE IN BOLD. ASTERISKS MARK SIGNIFICANT DIFFERENCES FROM OUR METHOD, AS PER A PAIRED STUDENT'S T-TEST. (*: P<0.05).

Method	Dice (%) at various foreground-background ratios					
	0%∼2%	2%∼4%	4%∼6%	6%∼8%	8%∼10%	Avg
ResU-Net [21]	54.76±6.03	73.69±5.27	74.99±5.36	75.21±5.03	75.26±4.73	62.61±6.54
nnU-Net [4]	62.79±4.73*	83.51±4.79*	**84.99±4.19**	**85.29±4.95**	85.23±4.57*	71.40±5.63*
Swin-Unet [14]	60.33±5.09	82.71±5.09	83.92±4.59	84.19±5.01	84.22±4.79	69.44±6.31
TransUNet [24]	57.93±6.18	79.55±5.76	81.21±6.39	80.77±5.96	79.58±6.10	66.82±6.22
TransFuse-S [15]	55.79±6.23	78.69±6.14	79.65±5.61	80.09±5.44	80.05±5.73	65.18±6.71
RSTN [16]	57.68±5.26	79.96±4.79	81.33±4.46	83.52±4.59	84.09±4.07	67.10±5.16
UFL [41]	58.12±5.40	78.79±5.63	80.19±6.17	80.88±5.20	81.76±5.03	66.76±5.41
Ours	**64.53±3.65**	**84.79±4.12**	84.73±4.51	85.21±4.03	**85.79±4.22**	**72.68±4.27**

Method	HD at various foreground-background ratios					
	0%∼2%	2%∼4%	4%∼6%	6%∼8%	8%∼10%	Avg
ResU-Net [21]	19.85±6.69	18.23±6.27	18.21±6.53	17.41±5.20	17.06±5.17	19.10±6.95
nnU-Net [4]	14.95±5.36*	15.22±6.23*	14.27±5.35	**13.59±4.36**	13.97±4.55*	14.80±4.96*
Swin-Unet [14]	15.76±5.23	16.17±6.71	15.04±5.13	14.21±5.63	14.39±4.78	15.62±6.13
TransUNet [24]	16.33±6.17	17.29±6.63	18.29±6.77	16.57±5.98	15.96±5.42	16.72±6.66
TransFuse-S [15]	16.76±6.23	16.97±5.96	17.75±6.29	17.77±5.19	16.53±4.92	16.96±6.74
RSTN [16]	17.10±5.29	16.26±6.71	**14.01±5.33**	14.93±6.12	14.71±5.41	16.47±5.63
UFL [41]	16.93±6.10	16.79±5.23	15.44±6.12	15.81±5.09	14.73±4.96	16.60±5.77
Ours	**14.31±5.22**	**14.79±6.19**	14.17±5.03	13.96±4.52	**13.27±4.73**	**14.26±5.11**

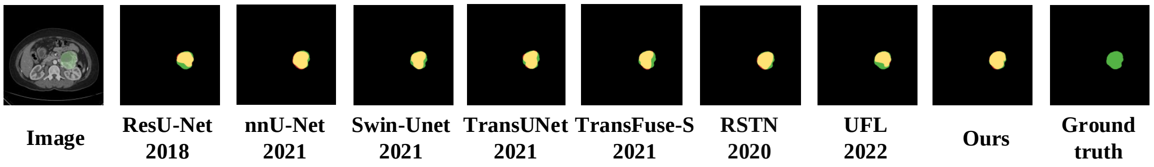

Fig. 6. Visual comparison of lesion segmentation results obtained from various methods on CT images from JLUKT. Red, green, and yellow indicate the prediction, ground truth, and overlapped pixels, respectively.

First, our MRL-Seg outperforms other methods in segmenting foreground and background for almost all foreground-background ratios, except the 4%-8% interval. These findings highlight the effectiveness of our method in handling imbalanced segmentation tasks, particularly in scenarios where the foreground-background ratio is highly imbalanced. Second, regarding stability, our approach exhibits superior performance in Dice and HD, showing the effectiveness of our method in integrating prior knowledge. Fig. 5 shows the qualitative results of our method. In comparison to other methods, our MRL-Seg achieves the most satisfactory predictions at extreme foreground-background ratios, with high agreement with the quantitative results. This demonstrates that our proposed method exhibits greater robustness compared to other methods, making it more suitable for imbalanced lesion segmentation. In addition, the p-value also derived from the Student's T-test further underscores the superiority of our approach.

Table II shows the quantitative results of our method on the JLUKT dataset, and Fig. 6 shows the corresponding visual comparison. Similar to the segmentation of kidney cancer in the KiTS19 dataset, we observed an improved model performance for almost all foreground-background ratios, except the 4%-8% ratio interval. As the foreground-background ratio was gradually balanced, the advantage became less pronounced. Our proposed method demonstrated superior performance compared to all state-of-the-art methods, as shown in Fig. 6, exhibiting contours that closely resemble real anatomical structures.

The quantitative results on the BUSI dataset are presented in Table III. The average Dice of our MRL-Seg model has reached 78.79%, surpassing the state-of-the-art Transformer-based model [14] [24] [15]. Furthermore, with the RL training strategy, performance improves even in the most challenging scenarios. It should be highlighted that the observed large variance can be traced back to the combination of complex and straightforward cases. The illustration of the segmentation result is depicted in Fig. 7. This visually demonstrates the effectiveness of our MRL-Seg, particularly in scenarios with extreme foreground-background ratios.

Table IV presents a comparative analysis of our proposed method with the state-of-the-art techniques on the NIH Pancreas Dataset. Compared to the other three datasets, the NIH dataset exhibits a smaller disparity in the size of the foreground elements (0%-2.3%). Consequently, there is no necessity to partition the foreground-background ratio into multiple intervals in our analysis. Our approach consistently outperforms the comparative methods in the Dice score and the Hausdorff distance index. This demonstrates that our approach possesses significant generalization capability in segmenting small objects, particularly in scenarios where there is an imbalance between the foreground and background. The illustration of the segmentation result is depicted in Fig. 8.

TABLE III
THE SEGMENTATION RESULTS (MEAN ± STD) FROM STATE-OF-THE-ART METHODS ON THE BUSI DATASET ARE SHOWN. THE BEST RESULTS ARE IN BOLD. ASTERISKS MARK SIGNIFICANT DIFFERENCES FROM OUR METHOD, AS PER A PAIRED STUDENT'S T-TEST. (*: P<0.05).

Method	Dice (%) at various foreground-background ratios						
	0%∼10%	10%∼20%	20%∼30%	30%∼40%	40%∼50%	50%∼60%	Avg
ResU-Net [21]	67.21±3.94	73.55±2.11	83.27±2.29	83.92±2.69	69.85±1.26	87.11±1.98	72.53±2.51
nnU-Net [4]	73.47±2.09*	78.79±1.13	85.71±1.10	**85.27±1.09**	71.22±1.19	88.77±0.54	77.45±1.13*
Swin-Unet [14]	70.23±2.11	76.55±1.47	86.21±1.03	81.19±1.79	66.29±2.21	86.53±0.71	74.81±2.17
TransUNet [24]	71.35±3.46	78.70±2.21	84.42±1.15	83.57±2.18	65.74±1.92	88.86±1.38*	75.94±1.84
TransFuse-S [15]	70.12±2.65	77.44±1.06	84.25±2.29	81.17±1.96	66.94±2.87	85.31±1.16	74.69±1.77
RSTN [16]	69.21±2.12	79.11±1.55*	**86.29±2.56**	84.19±1.99	73.84±2.47*	88.52±1.95	75.42±1.41
UFL [41]	69.19±3.54	76.51±1.49	84.23±2.95	84.36±1.96	70.09±1.56	86.15±1.33	74.29±1.12
Ours	**75.21±2.11**	**80.02±2.14**	85.93±1.20	84.74±1.21	**74.13±1.22**	**89.51±0.87**	**78.79±1.19**

Method	HD at various foreground-background ratios						
	0%∼10%	10%∼20%	20%∼30%	30%∼40%	40%∼50%	50%∼60%	Avg
ResU-Net [21]	30.98±3.79	33.52±2.54	25.47±3.21	24.59±1.67	34.96±3.78	23.95±3.29	30.37±2.66
nnU-Net [4]	25.21±2.64*	23.96±2.17*	**20.11±2.29**	22.51±2.62	30.21±1.77*	20.51±2.66	24.12±2.07*
Swin-Unet [14]	28.96±3.64	28.57±2.03	24.00±2.32	**21.24±2.20**	34.25±1.11	22.51±2.56	27.78±2.39
TransUNet [24]	27.15±3.30	29.55±1.13	22.34±1.23	23.44±1.68	33.24±2.58	21.31±1.29	27.97±2.69
TransFuse-S [15]	29.21±3.55	29.78±2.63	21.23±2.10	22.14±2.96	34.09±3.61	20.09±2.33*	27.74±3.26
RSTN [16]	26.93±2.36	24.22±1.57	21.26±1.22	23.91±2.77	30.29±2.49	21.03±2.56	25.29±2.41
UFL [41]	25.23±3.36	29.21±2.00	23.23±2.41	25.20±2.78	32.26±2.84	20.36±2.27	25.82±2.73
Ours	**22.03±2.11**	**23.31±2.03**	20.49±1.30	21.32±1.91	**29.44±1.26**	**19.85±1.10**	**21.20±2.03**

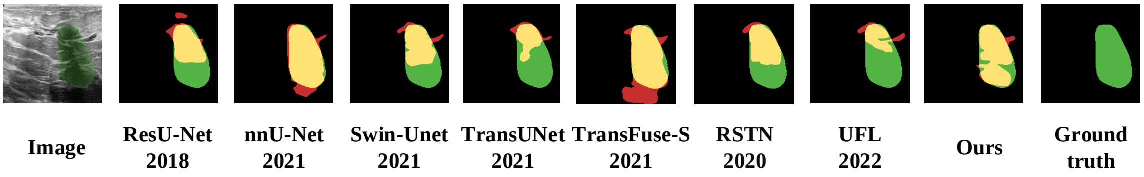

Fig. 7. Visual comparison of lesion segmentation results obtained from various methods on CT images from BUSI. Red, green, and yellow indicate the prediction, ground truth, and overlapped pixels, respectively.

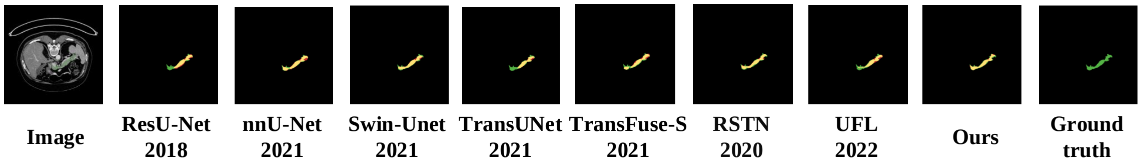

Fig. 8. Visual comparison of lesion segmentation results obtained from various methods on CT images from NIH. Red, green, and yellow indicate the prediction, ground truth, and overlapped pixels, respectively.

TABLE IV
THE SEGMENTATION RESULTS (MEAN ± STD) FROM STATE-OF-THE-ART METHODS ON THE NIH DATASET ARE SHOWN. THE BEST RESULTS ARE IN BOLD. ASTERISKS MARK SIGNIFICANT DIFFERENCES FROM OUR METHOD, AS PER A PAIRED STUDENT'S T-TEST. (*: P<0.05).

Method	Dice (%)	HD
ResUNet [21]	74.25±4.77	13.96±4.69
nnUNet [4]	81.21±3.12	11.08±4.22
SwinUNet [14]	77.97±4.10	12.71±4.57
TransUNet [24]	77.03±4.20	13.44±4.72
TransFuse [15]	76.28±4.75	12.01±4.96
RSTN [16]	81.73±3.52*	10.97±4.10*
UFL [41]	76.42±4.17	11.41±4.36
Ours	**82.29±2.51**	**10.13±3.48**

E. Ablation Analysis

We further explore the impact of the various components in our proposed model. Table V shows the comparison results of five variants, including: (1) ”OnlySeg” that utilizes only the ”Segmentation-Net” in Fig. 4; (2) ”RL+Seg” which corresponds to the proposed reinforcement learning-based segmentation framework in Fig. 2; (3) ”RL+FB+Seg” that combines the ”FB” module (Section III.B) and the proposed RL segmentation framework; (4) ”RL+FB+ADP+Seg” that further employs ”FB” and ”ADP” (Section III.C) during the RL segmentation framework training; (5) ”w/o Pre-training” that corresponds to the ”RL+Seg” which training from scratch without utilizing the parameters of ”Segmentation-Net”. In order to validate the effectiveness of these components, we conducted a set of ablation experiments.

As shown in Table V, the most decline in performance occurs when pre-training is not utilized. This reveals that training from scratch is hard. The pre-training as a warm start is undoubtedly crucial for reinforcement learning. Moreover, the performance of ablated version without FB drops significantly since the constraint of the foreground-background ratio alleviates the class-imbalance caused by ”input imbalance”.

TABLE V
ABLATION ANALYSIS OF KEY COMPONENTS

Model					Mean Dice (%)			
w/o pre-training	Seg	RL	ADP	FB	KiTS19	JLUKT	BUSI	NIH
✓	✓	✓			19.41±10.99	22.71±11.25	32.06±6.33	47.12±7.16
	✓				61.22±6.79	65.73±6.51	74.33±2.32	76.51±4.37
	✓	✓			65.29±6.01	69.15±5.76	77.23±2.59	80.70±3.87
	✓	✓	✓		67.41±5.13	70.93±5.04	77.93±2.03	81.55±3.19
	✓	✓	✓	✓	**68.94±4.19**	**72.68±4.27**	**78.79±1.19**	**82.29±2.51**
Model					**Mean HD**			
w/o pre-training	Seg	RL	ADP	FB	KiTS19	JLUKT	BUSI	NIH
✓	✓	✓			31.54±14.24	29.87±15.16	30.27±7.90	28.13±14.56
	✓				18.97±7.22	17.41±6.53	28.01±2.85	14.13± 4.74
	✓	✓			16.55±6.54	15.83±6.29	24.63±2.22	12.19±4.27
	✓	✓	✓		15.31±5.83	14.68±5.71	23.17±2.19	11.13±3.55
	✓	✓	✓	✓	**15.20±5.96**	**14.26±5.11**	**21.20±2.03**	**10.13±3.48**

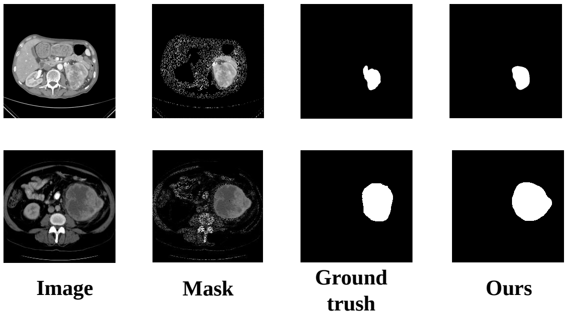

Fig. 9. Visualization of the masking results of patches learned by reinforcement learning.

In addition, the performance of the ablated version without ADP also demonstrates a significant decline, because ADP supplies local rewards to the reinforcement learning framework within our model. The integration of these modules into our proposed MRL-Seg model results in a notable enhancement in segmentation performance. This indicates that these two types of constraints work synergistically to enhance the model's adaptability, resulting in improved precision across nearly all substructures.

We observe the alterations in the mask of source images, as depicted in Fig. 9. Our model refrains from masking a patch that was initially completely black in the original image, implying that the black patch does not influence the segmentation process. It is evident from our observations that the model is proficient in discerning the non-ROIs within the organ tissue, subsequently applying dense mask annotations with ease. Moreover, the contrast of the ROIs in the masked image is notably enhanced compared to that in the original image. Interestingly, even with a few patches masked from the lesion-containing portion of the original image, an improved prediction image can still be segmented. These examples show that a purified representation can benefit the image segmentation task.

V. Conclusion

In this paper, we proposed a novel MRL-Seg for highly imbalanced medical image segmentation, which utilizes multi-step reinforcement learning. MRL-Seg comprises three modules: Reinforcement-Net, Representation-Net, and Segmentation-Net, which collectively model the significance of medical image patches. Our proposed method has several distinct advantages. Firstly, the Reinforcement-Net based on multi-step reinforcement learning effectively mitigates the negative impact of imbalanced data bias, while enabling the Representation-Net to learn the importance of different pixels for segmentation during the training process. Additionally, the Reinforcement-Net leverages 2D priors to assist in segmenting foreground and background, providing explicit guidance to complement the Segmentation-Net and leading to improved segmentation stability. Extensive experimental results showed that our proposed MRL-Seg could significantly improve the segmentation performance of highly imbalanced medical images and outperformed the state-of-the-art methods, which is of great significance for clinical diagnosis and research.

Acknowledgments

This work was supported by the National Natural Science Foundation of China [grant number 61801190]; the National Key Research and Development Project of China [grant number 2019YFC0409105]; the Thirteenth Five-Year Plan Scientific Research Planning Project of the Education Department of Jilin Province [grant numbers JJKH20200997KJ, JJKH20200678KJ]; and the Science and Technology Development Plan Project of Jilin Province[grant number 20230508164RC].

REFERENCES

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[2] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes," *IEEE transactions on medical imaging*, vol. 37, no. 12, pp. 2663–2674, 2018.

[3] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians*, vol. 71, no. 3, pp. 209–249, 2021.

[4] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.

[5] S. A. Taghanaki, Y. Zheng, S. K. Zhou, B. Georgescu, P. Sharma, D. Xu, D. Comaniciu, and G. Hamarneh, "Combo loss: Handling input and output imbalance in multi-organ segmentation," *Computer Vision and Pattern Recognition*, 2018.

[6] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep learning in medical image analysis and multimodal learning for clinical decision support*. Springer, 2018, pp. 3–11.

[7] Z. Wang, X. Li, H. Duan, Y. Su, X. Zhang, and X. Guan, "Medical image fusion based on convolutional neural networks and non-subsampled contourlet transform," *Expert Systems with Applications*, vol. 171, p. 114574, 2021.

[8] Z. Yan, X. Yang, and K.-T. Cheng, "A three-stage deep learning model for accurate retinal vessel segmentation," *IEEE journal of Biomedical and Health Informatics*, vol. 23, no. 4, pp. 1427–1436, 2018.

[9] H. Kervadec, J. Bouchtiba, C. Desrosiers, E. Granger, J. Dolz, and I. B. Ayed, "Boundary loss for highly unbalanced segmentation," in *International conference on medical imaging with deep learning*. PMLR, 2019, pp. 285–296.

[10] G. Wang, M. A. Zuluaga, W. Li, R. Pratt, P. A. Patel, M. Aertsen, T. Doel, A. L. David, J. Deprest, S. Ourselin *et al.*, "Deepigeos: a deep interactive geodesic framework for medical image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 7, pp. 1559–1572, 2018.

[11] M. S. Alam, D. Wang, Q. Liao, and A. Sowmya, "A multi-scale context aware attention model for medical image segmentation," *IEEE Journal of Biomedical and Health Informatics*, 2022.

[12] Y. Yuan, L. Zhang, L. Wang, and H. Huang, "Multi-level attention network for retinal vessel segmentation," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 1, pp. 312–323, 2021.

[13] Z. Guo, L. Zhao, J. Yuan, and H. Yu, "Msanet: Multiscale aggregation network integrating spatial and channel information for lung nodule detection," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 6, pp. 2547–2558, 2021.

[14] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," *arXiv preprint arXiv:2105.05537*, 2021.

[15] Y. Zhang, H. Liu, and Q. Hu, "Transfuse: Fusing transformers and cnns for medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 14–24.

[16] 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.

[17] S. Zhang, B. Ren, Z. Yu, H. Yang, X. Han, X. Chen, Y. Zhou, D. Shen, and X.-Y. Zhang, "Tw-net: Transformer weighted network for neonatal brain mri segmentation," *IEEE Journal of Biomedical and Health Informatics*, 2022.

[18] S. Di, Y.-Q. Zhao, M. Liao, F. Zhang, and X. Li, "Td-net: A hybrid end-to-end network for automatic liver tumor segmentation from ct images," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 3, pp. 1163–1172, 2022.

[19] 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.

[20] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio, "The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 11–19.

[21] M. Z. Alom, M. Hasan, C. Yakopcic, T. M. Taha, and V. K. Asari, "Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation," *arXiv preprint arXiv:1802.06955*, 2018.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[23] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[24] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[25] L. Xie, W. Cai, and Y. Gao, "Dmcgnet: A novel network for medical image segmentation with dense self-mimic and channel grouping mechanism," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 10, pp. 5013–5024, 2022.

[26] H. Li, J. Fang, S. Liu, X. Liang, X. Yang, Z. Mai, M. T. Van, T. Wang, Z. Chen, and D. Ni, "Cr-unet: A composite network for ovary and follicle segmentation in ultrasound images," *IEEE journal of biomedical and health informatics*, vol. 24, no. 4, pp. 974–983, 2019.

[27] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel, "Medical transformer: Gated axial-attention for medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 36–46.

[28] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[29] M. Moravčík, M. Schmid, N. Burch, V. Lisỳ, D. Morrill, N. Bard, T. Davis, K. Waugh, M. Johanson, and M. Bowling, "Deepstack: Expert-level artificial intelligence in heads-up no-limit poker," *Science*, vol. 356, no. 6337, pp. 508–513, 2017.

[30] D. Li, H. Wu, J. Zhang, and K. Huang, "A2-rl: Aesthetics aware reinforcement learning for image cropping," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8193–8201.

[31] J. Park, J.-Y. Lee, D. Yoo, and I. S. Kweon, "Distort-and-recover: Color enhancement using deep reinforcement learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5928–5936.

[32] R. Furuta, N. Inoue, and T. Yamasaki, "Fully convolutional network with multi-step reinforcement learning for image processing," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 3598–3605.

[33] X. Liao, W. Li, Q. Xu, X. Wang, B. Jin, X. Zhang, Y. Wang, and Y. Zhang, "Iteratively-refined interactive 3d medical image segmentation with multi-agent reinforcement learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9394–9402.

[34] G. Tao, H. Li, J. Huang, C. Han, J. Chen, G. Ruan, W. Huang, Y. Hu, T. Dan, B. Zhang *et al.*, "Seqseg: A sequential method to achieve nasopharyngeal carcinoma segmentation free from background dominance," *Medical Image Analysis*, vol. 78, p. 102381, 2022.

[35] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," in *International conference on machine learning*. PMLR, 2016, pp. 1928–1937.

[36] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM computing surveys (csur)*, vol. 53, no. 3, pp. 1–34, 2020.

[37] S. Yu, X. Li, Y. Feng, X. Zhang, and S. Chen, "An instance-oriented performance measure for classification," *Information Sciences*, vol. 580, pp. 598–619, 2021.

[38] N. Heller, F. Isensee, K. H. Maier-Hein, X. Hou, C. Xie, F. Li, Y. Nan, G. Mu, Z. Lin, M. Han *et al.*, "The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge," *Medical Image Analysis*, p. 101821, 2020.

[39] H. R. Roth, L. Lu, A. Farag, H.-C. Shin, J. Liu, E. B. Turkbey, and R. M. Summers, "Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part I 18*. Springer, 2015, pp. 556–564.

[40] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, "Dataset of breast ultrasound images," *Data in brief*, vol. 28, p. 104863, 2020.

[41] M. Yeung, E. Sala, C.-B. Schönlieb, and L. Rundo, "Unified focal loss: Generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation," *Computerized Medical Imaging and Graphics*, vol. 95, p. 102026, 2022.