Self-Supervised Medical Image Segmentation Using Deep Reinforced Adaptive Masking

Zhenghua Xu, *Member, IEEE*, Yunxin Liu, Gang Xu, and Thomas Lukasiewicz

***Abstract*— Self-supervised learning aims to learn transferable representations from unlabeled data for downstream tasks. Inspired by masked language modeling in natural language processing, masked image modeling (MIM) has achieved certain success in the field of computer vision, but its effectiveness in medical images remains unsatisfactory. This is mainly due to the high redundancy and small discriminative regions in medical images compared to natural images. Therefore, this paper proposes an adaptive hard masking (AHM) approach based on deep reinforcement learning to expand the application of MIM in medical images. Unlike predefined random masks, AHM uses an asynchronous advantage actor-critic (A3C) model to predict reconstruction loss for each patch, enabling the model to learn where masking is valuable. By optimizing the non-differentiable sampling process using reinforcement learning, AHM enhances the understanding of key regions, thereby improving downstream task performance. Experimental results on two medical image datasets demonstrate that AHM outperforms state-of-the-art methods. Additional experiments under various settings validate the effectiveness of AHM in constructing masked images.**

***Index Terms*— Self-supervised learning, medical image segmentation, adaptive image masking, deep reinforcement learning.**

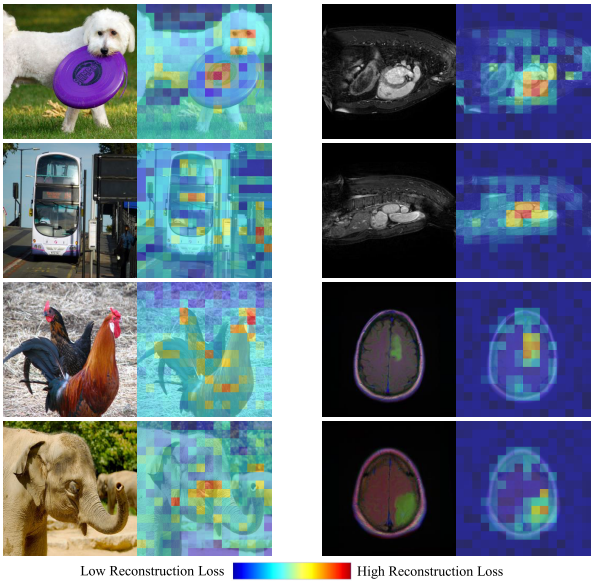


Fig. 1. Comparing natural and medical images in terms of content. For each tuple, we show the image (left), and the reconstruction loss predicted for each patch (right).

I. Introduction

THE goal of self-supervised learning (SSL) [1] is to learn transferable representations from a large amount of unlabeled data and apply them to downstream tasks, such as classification and segmentation. Inspired by masked language modeling (MLM) [2], [3], [4], [5] in natural language processing (NLP), masked image modeling (MIM) leverages the fact that natural images contain more redundant information compared to natural language, thus, many methods [2], [6], [7], [8] have applied this approach in computer vision (CV) domain and achieved promising results. However, if these MIM methods are directly applied to medical images, can they also achieve good performance?

In fact, for medical image data, the annotation cost is much higher than that of natural images and there is a large amount of unlabeled data. This means that SSL plays a more crucial role in the medical domain to reduce the need for costly annotations. For this purpose, we try to answer the above question. First, the information distribution in natural images is not uniform, i.e., some areas (with higher reconstruction loss) contain more information than others (with lower reconstruction loss), so having challenging masking strategies (i.e., tend to mask areas with higher reconstruction loss) is crucial [9]; consequently, some traditional MIM pretraining methods [2], [6], [7] are proposed to generate such strategies in a predefined manner. However, the information distribution in medical images is more uneven compared to natural images; as shown in Figure 1, information of natural images is usually highly dispersed and diverse compared with medical images, and medical images contains much less high

Manuscript received 9 June 2024; accepted 24 July 2024. Date of publication 1 August 2024; date of current version 2 January 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62276089; in part by the Natural Science Foundation of Hebei Province, China, under Grant F2024202064; in part by the Ministry of Human Resources and Social Security, China, under Grant RSTH-2023-135-1; and in part by the AXA Research Fund. *(Zhenghua Xu and Yunxin Liu are co-first authors.) (Corresponding author: Zhenghua Xu.)*

Zhenghua Xu and Yunxin Liu are with State Key Laboratory of Reliability and Intelligence of Electrical Equipment, School of Health Sciences and Biomedical Engineering, Hebei University of Technology, Tianjin 300401, China (e-mail: zhenghua.xu@hebut.edu.cn; liudearbreeze@gmail.com).

Gang Xu is with the School of Artificial Intelligence, Hebei University of Technology, Tianjin 300401, China (e-mail: 1710822388@qq.com).

Thomas Lukasiewicz is with the Department of Computer Science, University of Oxford, OX1 3QG Oxford, U.K., and also with the Institute of Logic and Computation, Vienna University of Technology, 1040 Vienna, Austria (e-mail: thomas.lukasiewicz@cs.ox.ac.uk).

Digital Object Identifier 10.1109/TMI.2024.3436608

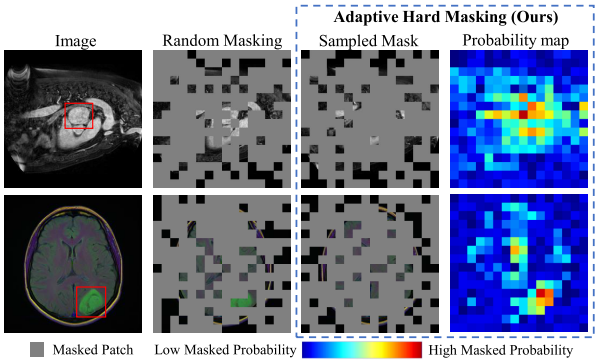


Fig. 2. Comparison of our adaptive masking with existing random patch masking [6], [7], [8] for masking ratio of 75%. Our adaptive masking approach (third column) selects more masked patches from the regions where the model predicts to have high reconstruction loss (fourth column) to produce a suitable and challenging pretext task.

construction loss areas than nature images; therefore, using the predefined random mask strategy as in natural images may not be sufficient to obtain satisfactory medical image segmentation performances. As shown in the second column of Figure 2, where some key organs/lesions are mostly selected as visible patches, which leads to less learned organ/lesion representation information and limit downstream task performance. In addition, a high masking rate means that the learnable conditional mutual information is small and the gradient noise is large, which limits the upper limit of the representation information learned by the network. Therefore, learning a suitable and challenging mask task in medical images is crucial. As shown in the third column of Figure 2, we hope that for different medical images, the model can learn to create suitable and challenging problems to help the reconstruction network improve the understanding of key regions, thereby maximizing the release of the performance limitations of downstream tasks.

To this end, we propose a new MIM training paradigm called ***A**dapted **H**ard **M**asking (AHM)* based on deep reinforcement learning for medical image segmentation tasks. Specifically, given an input image, we do not generate binary masks according to a predefined uniform random sampling criterion. Instead, we use an asynchronous advantage actor-critic (A3C) [10]-based deep reinforcement learning model to predict the reconstruction loss for each patch. We then sample invisible blocks from this distribution, obtaining suitable and highly challenging masks, and then train the reconstruction network to predict masked patches just like conventional methods. Through this way, encouraging the model to learn where it is worth being masked. Since sampling is a non-differentiable operation, we optimize this process using an actor-critical reinforcement learning strategy. Additionally, our approach does not directly use the computed reconstruction loss as a target for guiding network parameter updates. The reconstruction loss is only used to calculate rewards, making it part of the auxiliary network parameter updates. Thus, our model is not overwhelmed by the precise values of the reconstruction loss. As shown in the last two elements of each tuple in Figure 2, patches with larger prediction losses tend to be discriminative. Therefore, masking these patches introduces a challenging situation where organ/lesion regions are almost entirely masked. However, as proved in [9], setting too difficult reconstruction targets, i.e., masking nearly all the areas with high Reconstruction losses, will also hinder the model's ability to learn appropriate masking strategies, so we additionally introduce a random sampling strategy to randomly select a certain proportion of patches to be retained (i.e., they are not allowed to be masked). This will ensure that the model will still retain some reasonable hints to learn the features of high reconstruction loss areas during training.

In summary, this work's main contributions are as follows:

- We find that the current masking image modeling method applied to medical images have limitations due to the high redundancy and small discriminative regions in medical images compared to natural images. Therefore, we propose a novel masking patch selection strategy AHM for medical image segmentation tasks.
- We model the process of masking images as a deep reinforcement learning problem, which uses a reconstruction network to provide feedback signals to guide the A3C-based deep reinforcement learning to learn image-specific masking strategies, selecting suitable and challenging masks for each image.
- Extensive experiments are conducted on two public medical image datasets. The experimental results show that the proposed AHM outperforms the state-of-the-art (SOTA) method by 2.8% and 1.6% (resp. 2.9% and 2.3%) on the DSC metric for the Cardiac and TCIA datasets with 5% labeled data (resp. 10% labeled data).

II. Related Work

A. Self-Supervised Learning

1) Contrastive Learning: It is primarily based on the construction of positive and negative samples, followed by the comparison of distances between them. Methods such as SimCLR [11] achieve this by applying random data augmentations to a batch of input images, aiming to maximize the similarity between positive samples while minimizing the similarity between negative samples. BYOL [12] takes a different approach by eliminating the need for negative samples and instead constraining the online network to predict the target network representation of the same image under various augmented versions, utilizing the mean square error (MSE) loss. SwAV [13] introduces clustering algorithms that enhance the consistency of cluster assignments across different views of the same image, facilitating the learning of informative representations. PCRL [14] introduce Preservational Contrastive Representation Learning to reconstruct diverse contexts using representations learned from the contrastive loss. However, these approaches tend to emphasize the acquisition of global semantic features in images, which may limit their capability to capture fine-grained representations. Consequently, their effectiveness in downstream segmentation tasks may be compromised. In contrast, the proposed AHM overcomes this limitation and enables pre-trained models to

learn finer-grained image features, thereby leading to improved accuracy in downstream segmentation models.

2) Masked Image Modeling: It is a technique that focuses on learning fine-grained visual representations by reconstructing the masked regions within an image. In the work by Deepak et al. [15], a fixed central area of an image is masked, and the network leverages the surrounding image information to infer the missing region. MG [16] uses a unified self-supervised approach built directly from unlabeled 3D image data for generating powerful application-specific target models through transfer learning. MAE [6] adopts an asymmetric encoder and decoder structure, dividing the image into equally-sized patches, and predicts the masked patches based on the unmasked image patches. SimMIM [7] builds upon MAE by adjusting the decoder's weight and taking both visible and masked patches as input, achieving similar results as MAE while accelerating the pre-training process. ConvMAE [8] employs multi-scale coding operations based on MAE, enabling the model to learn richer semantic information. ADIOS [17] proposed adversarial masking and adversarial discriminative instance optimization for Self-Supervised Learning to improve self-supervised learning by enhancing the quality and diversity of learned representations. AttMask [18] generates an attention map through teacher transformer encoder, which it use to guide masking for the student encoder. However, in these methods, the positions and sizes of the masks are either fixed or randomly determined, or the mask strategy is based on natural images and cannot be well adapted to medical images, lacking adaptability to individual images, which leads to suboptimal pre-trained model weights. In contrast, AHM learns the masking policy is crucial. This enables the reconstruction network to be guided in a more challenging manner, allowing it to learn the most informative parts of the image and improve the performance of downstream tasks.

B. Deep Reinforcement Learning

Inspired by the successful application of deep reinforcement learning in playing video games [19], many deep reinforcement learning algorithms applied in medical image analysis [20], especially medical image segmentation [21], have been widely explored and applied. Wang et al. [22] propose an online reinforcement learning framework for medical image segmentation, which first introduces the concept of context-specific segmentation such that the model is adaptive not only to a defined objective function but also to the user's intention and prior knowledge. Liao et al. [23] propose multi-agent reinforcement learning with user interaction to capture the dependency among voxels for medical image segmentation and to reduce the exploration space to a tractable size. Tian et al. [24] propose an end-to-end deep reinforcement learning to mimic physicians delineating a region of interest (ROI) on the medical image in a multi-step manner from a coarse result to a fine result progressively. Man et al. [25] apply deep Q-learning (DQN) to identify each pancreas's bounding box and then use a modified U-Net to segment the pancreas in cropped CT images. Qin et al. [26] propose to train both augmentation and segmentation modules simultaneously and use the errors during segmentation as feedback to adjust the augmentation module. DRL-LNS [27] proposes a DRL method for weakly supervised lesion segmentation. However, these DRL methods are based on fully supervised or weak-supervised learning, and thus they cannot utilize unlabeled data. Our proposed AHM is based on SSL, which effectively uses unlabeled data to pre-train the segmentation network, reducing the labeling cost while ensuring segmentation accuracy, and effectively solving the problem of scarcity of medical image labeled data.

TABLE I
FREQUENTLY USED SYMBOLS

Symbol	Explanation
s	Current State
a	Action
r	Reward
γ	The discount factor
θ_v	Parameters of value network
θ_p	Parameters of policy network
θ_f	Parameters of feature extraction network
$A(a,s)$	Advantage function
$\pi(a\|s)$	Policy function
$V(s)$	State value function
$P \times P$	The number of patches

III. BACKGROUND KNOWLEDGE

In this paper, we propose an extension of the asynchronous advantage actor-critic (A3C) [10] algorithm to address the problems of current medical image masking strategies, as A3C has demonstrated effective performance and efficient training in the original paper. In this section, we provide a brief overview of the training algorithm employed by A3C. A3C belongs to the actor-critic family of methods and consists of two networks: the policy network and the value network. The policy network encourages the agent to make better actions, and the value network scores more accurately based on status. We denote the parameters of each network as θ_p and θ_v, respectively. Both networks take the current state $s^{(t)}$ as input. The value network produces the value estimate $V(s^{(t)})$, which represents the expected total rewards from the current state and indicates the quality of the state. Here, t represents discrete time step. Since our method does not require iteration, we omit t in all formulas for the convenience of understanding (i.e., $t = 1$). For the value network, we hope that its score will be closer to reward r. Therefore, the gradient for updating the parameters of the value network θ_v is computed as follows:

$$d\theta_v = \nabla_{\theta_v}(r - V(s))^2, \tag{1}$$

The policy network outputs the policy $\pi(a|s)$ of taking the action $a \in \delta$. so the output dimension of the policy network is $|\delta|$. For the policy network, we hope that it can make actions with higher reward, Therefore, gradient for updating the parameters of the policy networkθ_p is computed as follows:

$$A(a,s) = r - V(s), \tag{2}$$

$$d\theta_p = -\nabla_{\theta_p} \log \pi(a|s) A(a,s). \tag{3}$$

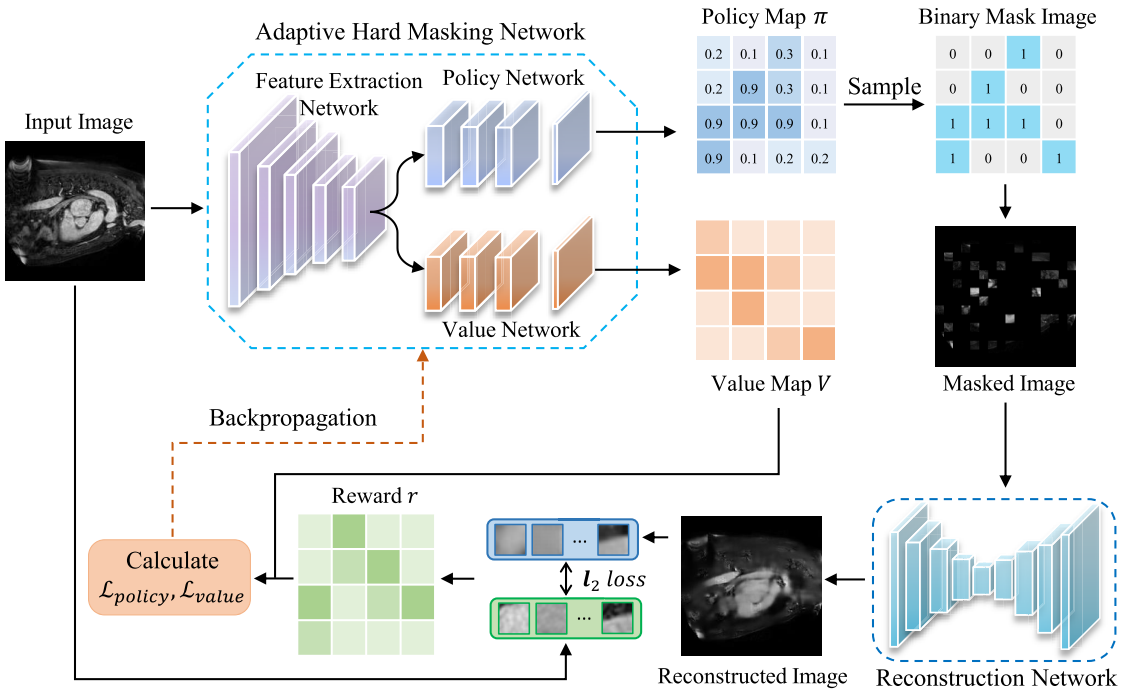


Fig. 3. The framework of the proposed AHM. Our AHM Architecture is based on A3C, where the adaptive hard masking network is used to predict regions with higher reconstruction loss. Our adaptive sampling procedure samples more patches from more discriminative regions and fewer patches from the low information or redundant regions.

where $A(a, s)$ is called the advantage, which represents the advantage of taking an action in the current state relative to the average expectation. And $V(s)$ is subtracted to reduce the variance of the gradient. To make the definitions of various symbols clearer, we have created Table I to explain the meanings of commonly used symbols.

Traditional actor-critic methods are primarily designed for games, and they are rarely used in image processing, especially self-supervised image processing tasks. Therefore, although we adopt the intuition and mechanism of original A3M, the A3C used in this work has been greatly redesigned to adopt with the task of self-supervised medical image processing to properly generate medical image masking policy.

Specifically, we first design a reasonable new reward function based on reconstruction loss to assist the algorithm in converging and encouraging it to identify areas with high reconstruction loss. Second, we have defined a new action according to the task requirements (i.e., whether to mask or not). Finally, we propose a new agent definition strategy to treat each patch block as an independent agent instead of treating game players as agents as in original A3C, helping the model to better determine which patches need to be masked and allowing them to make independent actions.

IV. Method

In view of the deficiencies of current masking image modeling methods in medical images, we believe that it is crucial to need suitable and challenging pretext tasks for medical images with greater redundancy. To address this, we introduce the Adaptive Hard Masking Network based on deep reinforcement learning to predict the reconstruction loss for each masked patch, and carefully design its optimization objective. Figure 3 provides an overview of our proposed AHM, which will be further explained below.

A. Architecture of Adaptive Hard Masking Network

Figure 3 shows our Adaptive Hard Masking Network architecture which consists of three main components: a feature extraction network (we use VGG16 [28] in this case, but theoretically, any feature extraction module can be used), a policy network, and a value network. The policy network and value network share the output of the feature extraction network. They have the same network architecture except for the last layer. The detailed network structure of the Adaptive Hard Masking Network is shown in Table III.

In each training iteration, the original image $X \in \mathbb{R}^{H\times W\times C}$ is first passed through the feature extraction network to extract the image feature information. The resulting feature maps have a resolution of $1/P$ of the original image, that is, the original image is divided into $P \times P$ non-overlapping patches. (H, W) is the resolution of the original image, C is the number of channels, $(H/P, W/P)$ is the size of each patch, and then feed it into the policy network and the value network respectively. The value network outputs the value $V(s) \in \mathbb{R}^{P\times P}$, where each pixel value corresponds to a patch. Each pixel value represents the state value of the respective patch. The output policy $\pi(a|s) \in \mathbb{R}^{P\times P}$ of the policy network has the same resolution as the input feature map. This is done to predict the reconstruction loss for each patch. Then, sampling is performed based on the predicted reconstruction loss to determine the action a, which is the Masked Patch. However, considering that the Mask image obtained purely in this way may be

too challenging for the reconstruction network, although the sampling operation has some randomness, as the number of training iterations increases, this randomness will decrease. We want to provide some reasonable hints to guide the reconstruction network in reconstructing the masked patches. Additionally, for reinforcement learning, a certain level of randomness can also help in finding potentially better policies. To this end, we reserve $p\%$ random sampling probability for each patch from the sampling process in $\pi(a|s)$, where p is an adjustable hyperparameter.

B. Optimizing Adaptive Hard Masking

Here, we describe the updated strategy of our method. We denote the i-th patch in the input image X divided into $P \times P$ patches ($i = 1, \ldots, P \times P$) as stats s_i. Each patch is an agent, and its policy is denoted as $\pi_i(a_i|s_i)$, where a_i and s_i are the actions and the state of the i-th agent, respectively. The agents obtain rewards $r = (r_1, \ldots, r_N)$ from the environment by taking the actions $a = (a_1, \ldots, a_N)$. The objective of the deep reinforcement learning problem is to learn the optimal policies $\pi = (\pi_1, \ldots, \pi_N)$ that maximize the mean of the total expected rewards at all patches:

$$\pi^* = \underset{\pi}{argmax} \;\; E_\pi(\overline{r}), \tag{4}$$

$$\overline{r}^{(t)} = \frac{1}{P \times P} r_i, \tag{5}$$

where r is the mean of the rewards r_i at all patches. The common solution is to divide the problem into $P \times P$ independent subproblems and train $P \times P$ networks, where we train the i-th agent to maximize the expected total reward at the i-th patch:

$$\pi_i^* = \underset{\pi_i}{argmax} \;\; E_{\pi_i}(r_i). \tag{6}$$

However, training $P \times P$ separate networks becomes impractical when the number of patches is large, and it only handles images with a fixed number of patches. To address these issues, we employ a convolutional network instead of $P \times P$ individual networks [29]. By using a convolutional network, all $P \times P$ agents can share parameters, and we can parallelize the computation of $P \times P$ agents on a GPU, making the training efficient, and since convolutional networks are used instead of independent multiple networks, each Agent is not isolated, our policy network and value network observe not only the i-th patch s_i but also the neighbor pixels to output the policy π and value V at the i-th patch. This method differs from typical deep reinforcement learning settings in two ways. First, this deep reinforcement learning approach involves an extremely large number of agents. Additionally, the agents are arranged in a 2D image plane. In this work, we use the update strategy from A3C to improve the performance of the agents.

In addition, we also need to design a good reward function to help guide the agent towards the desired goal. We believe that the ability to generate suitable and challenging masks is crucial for MIM pre-training. Intuitively, we think that hard patches are usually the patches with high reconstruction loss, which implicitly indicate the most discriminative part in the image, thus, we hope our model has the ability to predict the reconstruction loss of each patch. To this end, we use the loss function of the reconstruction network to calculate the reward, and the reward calculation formula is as follows:

$$r = \frac{1}{P \times P} \sum_{i=1}^{P \times P} \mathcal{L}_{rec}(X_i', X_i) \tag{7}$$

where X is the original image, X' is the reconstructed image output by the reconstruction network, and $\mathcal{L}_{rec}$ is ℓ_2-distance.

Since we divide the image into $P \times P$ patches and treat it as an independent agent, each agent has its own independent set of variables. For the convenience of understanding, we redefine Eqs. (1) to (3) in matrix form used to calculate multi-agent as follows:

$$d\theta_v = \nabla_{\theta_v} \frac{1}{N} \mathbf{1}^T \{(\mathbf{r} - \mathbf{V}(\mathbf{s})) \odot (\mathbf{r} - \mathbf{V}(\mathbf{s}))\}\mathbf{1}, \tag{8}$$

$$\mathbf{A}(\mathbf{a}, \mathbf{s}) = \mathbf{r} - \mathbf{V}(\mathbf{s}), \tag{9}$$

$$d\theta_p = -\nabla_{\theta_p} \frac{1}{N} \mathbf{1}^T \{\log \boldsymbol{\pi}(\mathbf{a}|\mathbf{s}) \odot \mathbf{A}(\mathbf{a}, \mathbf{s})\}\mathbf{1}, \tag{10}$$

where $\mathbf{r}$, $\mathbf{V}(s)$, $\mathbf{A}(\mathbf{a}, \mathbf{s})$ and $\boldsymbol{\pi}(\mathbf{a}|\mathbf{s})$ are the matrices whose (i_x, i_y)-th elements are r_i, $V(s_i)$, $A(a_i, s_i)$ and $\pi(a_i|s_i)$, respectively. $\mathbf{1}$ is an all-ones vector where every element is one, and $\odot$ means element-wise multiplication.

Similarly to the gradients of θ_p and θ_v, the gradient using the matrix form for the parameters of feature extraction network θ_f is computed as follows:

$$\begin{aligned} d\theta_f &= -\nabla_{\theta_f} \frac{1}{N} \sum_{i=1}^{N} \log \pi(a_i|s_i)(R_i - V(s_i)), \\ &\quad + \nabla_{\theta_f} \frac{1}{N} \sum_{i=1}^{N} (R_i - V(s_i))^2 \end{aligned} \tag{11}$$

$$\begin{aligned} &= -\nabla_{\theta_f} \frac{1}{N} \mathbf{1}^T \{\log \boldsymbol{\pi}(\mathbf{a}|\mathbf{s}) \odot \mathbf{A}(\mathbf{a}, \mathbf{s})\}\mathbf{1} \\ &\quad + \nabla_{\theta_f} \frac{1}{N} \mathbf{1}^T \{(\mathbf{r} - \mathbf{V}(\mathbf{s})) \odot (\mathbf{r} - \mathbf{V}(\mathbf{s}))\}\mathbf{1}. \end{aligned} \tag{12}$$

Similarly to typical policy gradient algorithms, the first term of $d\theta_f$ encourages a higher expected total reward. The second term operates as a regularizer such that R_i is not deviated from the prediction $V(s_i)$ by the convolution. We summarize the training algorithm of Adaptive Hard Masking in Algorithm 1.

V. Experiments

A. Datasets

The empirical studies over two real datasets confirm that our models beat other baseline models. As shown in Table II, we use two medical imaging datasets for model evaluation, which contain the characteristics of small datasets, small objects, and complex segmentation details (such as segmentation edges), and are more representative of the characteristics of current medical images.

Cardiac [30] is a public CT dataset to automatically segment the heart, which contains 20 cases. As the scanning mechanism is different, each case has 320×320 CT images with a range of 90 to 130 slices. The difficulty of its segmentation is that the dataset is small and the segmentation targets change greatly.

TCIA [31] is also a public MRI (fluid-attenuated inversion recovery sequence, FLAIR) dataset to automatically segment

Algorithm 1 Training Pseudo-Code of Adaptive Hard Masking

Input: Global counter T_{max}; training original images X; the pre-trained reconstruction network N; θ_p, θ_v, and θ_f for the policy network, value network and feature extraction network, respectively.

Assume global shared parameter vectors θ_p, θ_v, and θ_f
Assume thread-specific parameter vectors θ'_p, θ'_v, and θ'_f
for $T = 1, \ldots, T_{max}$ **do**
  Reset gradients: $d\theta_p \leftarrow 0$, $d\theta_v \leftarrow 0$, and $d\theta_f \leftarrow 0$
  Synchronize thread-specific parameters $\theta'_p = \theta_p$, $\theta'_v = \theta_v$, and $\theta'_f = \theta_f$
  Obtain state s
  Perform binary mask M according to policy $\pi(a|s; \theta'_p)$
  Obtain masked image X_m
  Input X_m to reconstruction network N to get reconstructed image X'
  Calculate reward r based on the reconstruction loss computed by X and X'
  $T \leftarrow T + 1$
  Accumulate gradients w.r.t. θ'_p: $d\theta_p \leftarrow d\theta_p - \nabla_{\theta'_p} \log \pi(a|s; \theta'_p)(r - V(s; \theta'_v))$
  Accumulate gradients w.r.t. θ'_v: $d\theta_v \leftarrow d\theta_v + \nabla_{\theta'_v}(r - V(s; \theta'_v))^2$
  Accumulate gradients w.r.t. θ'_f: $d\theta_f \leftarrow d\theta_f - \nabla_{\theta_f} \log \pi(a|s; \theta'_p)(r - V(s; \theta'_v)) + \nabla_{\theta_f}(r - V(s; \theta'_v))^2$
  Update θ_p, θ_v, and θ_f using $d\theta_p$, $d\theta_v$, and $d\theta_f$ respectively.
end for

Output: θ_p, θ_v, and θ_f

TABLE II
DATASETS INFORMATION

Datasets	Total	Normal	Abnormal	Size	Modality	Challenge
Cardiac [30]	2,271	921	1350	320×320	MRI	Small training dataset with large variability
TCIA [31]	3,929	2556	1373	256×256	MRI	Extremely irregular segmentation edges

TABLE III
NETWORK ARCHITECTURE OF ADAPTIVE HARD MASKING NETWORK, IN WHICH THERE ARE FOUR DOWNSAMPLING LAYERS IN THE FEATURE EXTRACTION NETWORK (I.E., $i = \mathbf{1}, \mathbf{2}, \mathbf{3}, \mathbf{4}$)

Feature Extraction Network		Input Shape	Output Shape
Input layer	Conv2d, BN2d, ReLU	$(1, h, w)$	(c, h, w)
Downsampling layer i	Conv2d, BN2d, ReLU	$(c, h, w/2^{i-1})$	$(c, h/2, w/2^i)$
Output layer	Conv2d, BN2d, ReLU	$(c, h/16, w/16)$	$(c, h/16, w/16)$

Policy Network	Input Shape	Output Shape	Value Network	Input Shape	Output Shape
Conv2d, BN2d, ReLU $\times 3$	$(c, h/16, w/16)$	$(c, h/16, w/16)$	Conv2d, BN2d, ReLU $\times 3$	$(c, h/16, w/16)$	$(c, h/16, w/16)$
Conv2d, Sigmoid	$(c, h/16, w/16)$	$(1, h/16, w/16)$	Conv2d	$(c, h/16, w/16)$	$(1, h/16, w/16)$

the tumor, which contains 110 cases. Segmentation masks for FLAIR abnormality are approved by a board-certified radiologist at Duke University. As the scanning mechanism is different, each case has 256×256 images with a number of slices ranging from 40 to 176. The segmentation challenge of this dataset is that the segmentation edges of objects have many complex segmentation details.

To train and evaluate the networks, all the above two medical image datasets are preprocessed as follows. First, for Cardiac, we transform these $3D$ images into $2D$ images according to the transverse section in a slice-by-slice manner [32]. Then, for both datasets, we normalize all input images to have zero mean and unit std. After that, to remove the images without the segmentation targets, negative samples are deleted from the datasets (the reasons of this operation are explained in Subsection V-G). In fact, there are $1,350$ images of Cardiac and $1,373$ images of TCIA used in our experiments. Finally, for all datasets, there are 70% of the datasets for training, 10% for validation, and 20% for testing.

B. Implementation Details

Our experiments are implemented using PyTorch[1] and run on an NVIDIA GeForce GTX 3090 GPU. To evaluate the performance of AHM, we perform fully supervised training on a randomly initialized U-Net using 5% and 10% of the data and use it as our original baseline. For the medical image segmentation task, we choose the U-Net structure as our reconstruction network, whose specific settings are consistent with the original paper [33]. The reconstruction network is pre-trained using all the training images with random masking.

[1]link: https://pytorch.org/

In the pre-training of the reconstructed network, we use the *Adam* [34] optimizer with the initial learning rate set to $2e-4$ and the batch size of 12. During the training process, the reconstructed network parameters is fixed. For transfer learning, the U-Net for downstream segmentation tasks is trained with the *Adam* optimizer with an initial learning rate set to $3e-4$, 10% decay every 3 epochs, and batch size of 4. For our adaptive hard masking network, we also use *Adam* optimizer with a learning rate set to $3e-4$, the learning rate drops by a factor of 0.9 every 25 epoch. The batch size is set to 4. We set the maximum epoch to 200. We present in Table III the structure and design of each layer of the adaptive hard masking network in detail. For all experiments, the number of divided patches is 256 (i.e., P is 16) and random sampling probability p is 30% unless otherwise noted.

C. Evaluation

To show the effectiveness of our models, we use the Dice Similarity Coefficient (DSC), Positive Predictive Value (PPV), Sensitivity (SEN), Intersection over Union (IoU), 95% Hausdorff Distance (HD95) [35] and boundary IoU (BIoU) [3]. Note that higher values for these metrics, except HD95, mean better performance. Formally,

$$DSC = \frac{2*TP+\epsilon}{T+P+\epsilon}, \quad PPV = \frac{TP+\epsilon}{TP+FP+\epsilon},$$
$$SEN = \frac{TP+\epsilon}{TP+FN+\epsilon}, \quad IoU = \frac{TP+\epsilon}{T+P-TP+\epsilon},$$
$$BIoU = \frac{G_d \cap P_d}{G_d \cup P_d},$$
$$HD95 = max_{k95\%}[d(P,G), d(G,P)]$$

where TP, FP, and FN are the number of true positive points, false positive points, and false negative points, respectively. T is the number of ground-truth points of that class, P is the number of predicted positive points, G is the number of ground-truth positive points, P_d is the number of predicted boundary positive points, G_d is the number of ground-truth positive boundary points, and $d(*)$ is a function to calculate surface distance. Finally, ϵ is a small constant to avoid zero division, which is set to $1e-4$ in our experiment.

D. Main Results

In order to evaluate the performances of the proposed AHM, randomly initialized U-Nets without self-supervised pretraining, i.e., fully supervised learning from scratch (denoted Fully Supervised), using 5% and 10% annotations are selected as our original baselines.

Several state-of-the-art self-supervised learning methods applied in the field of medical image segmentation are chosen as the self-supervised learning baselines in our experiments, namely, SimCLR [11], BYOL [12], SwAV [13], Context [15], PCRL [14], MG [16], SimMIM [7], MAE [6], ADIOS [17], ConvMAE [8], AttMask [18]. We evaluate the quality of the learned representations by transferring them from different self-supervised learning methods to the same downstream segmentation tasks. In order to evaluate the quality of the pre-trained model, we perform end-to-end fine-tuning (instead of linear probing). Then, we evaluate their impact on downstream performance. Finally, we also show the fully supervised results using large ratios (50% and 100%) of annotations.

To demonstrate the effectiveness of our proposed AHM, we conducted experiments on two public datasets and compare the performance of AHM with two state-of-the-art baselines: Fully Supervised Baseline (i.e., Fully Supervised), Self Supervised Baselines. The quantitative experimental results are shown in Table IV, examples of segmentation results of our AHM and baselines on the two datasets are shown in Figure 4. Through a quantitative and qualitative analysis, we obtain the following conclusions.

1) Compare With Fully Supervised Learning From Scratch: As shown in Table IV, AHM generally outperforms the baseline model trained from scratch by a large margin with 5% and 10% annotations. Specifically, in the case of 5% annotation, we first find that AHM is 38.97%, 52.90%, 16.41%, 40.04%, and 94.21% higher than the fully supervised method with 5% annotations on the Cardiac dataset for DSC, IoU, PPV, SEN, and BIoU, respectively, and HD95 is 3.3297 lower; while on the TCIA dataset, DSC, IoU, PPV, SEN, and BIoU are 19.74%, 22.03%, 15.24%, 6.35%, and 22.44% higher than the fully supervised method with 5% annotations, respectively, and HD95 is 0.4557 lower. And similarly, in the case of 10% annotation, AHM is 13.78%, 22.44%, 12.10%, 8.86%, and 37.16% higher than the fully supervised method with 10% annotations on the Cardiac dataset for DSC, IoU, PPV, SEN, and BIoU, respectively, and HD95 is 6.1151 lower; while on the TCIA dataset, DSC, IoU, PPV, SEN, and BIoU are 11.96%, 11.60%, 6.47%, 7.37%, and 26.54% higher than the fully supervised method with 10% annotations, respectively, and HD95 is 0.4342 lower. In addition, in the case of large ratios (50% and 100%) of annotations, AHM still outperforms the fully supervised model. Specifically, in the case of 50% annotation, AHM is 5.66%, 5.73%, 1.46%, 9.08%, and 11.74% higher than the fully supervised method with 50% annotations on the Cardiac dataset for DSC, IoU, PPV, SEN, and BIoU, respectively, and HD95 is 0.8208 lower; while on the TCIA dataset, DSC, IoU, PPV, and BIoU are 7.01%, 8.09%, 8.32%, and 2.09% higher than the fully supervised method with 50% annotations, respectively, and HD95 is 0.4924 lower. And similarly, in the case of 100% annotation, AHM is 1.72%, 2.61%, 2.43%, and 7.96% higher than the fully supervised method with 100% annotations on the Cardiac dataset for DSC, IoU, PPV, and BIoU, respectively, and HD95 is 0.3428 lower; while on the TCIA dataset, DSC, IoU, PPV, and BIoU are 2.99%, 3.89%, 7.65%, and 15.75% higher than the fully supervised method with 100% annotations, respectively, and HD95 is 0.0578 lower. Furthermore, the performance of AHM with 10% annotations can be close to or even better than that of the fully supervised method with 50% annotations. This is due to the ability of our method to learn more valuable representations from a large amount of unlabeled data, thereby improving the performance of downstream segmentation models.

2) Compare With Self-Supervised Learning Baselines: Then, we compare our AHM with the state-of-the-art selfsupervised methods on Cardiac and TCIA datasets with 5% and 10%

TABLE IV
RESULTS OF APPLYING THE PROPOSED METHOD AND THE STATE-OF-THE-ART BASELINES ON TWO PUBLIC DATASETS, WHERE THE BEST RESULTS ARE IN BOLD, THE SECOND BEST METHODS ARE UNDERLINED

Methods		Cardiac						TCIA					
		DSC↑	IoU↑	PPV↑	Sen↑	BIoU↑	HD95↓	DSC↑	IoU↑	PPV↑	Sen↑	BIoU↑	HD95↓
5%	U-Net	0.4836	0.3612	0.5873	0.5299	0.1505	12.1369	0.5763	0.5472	0.7785	0.7516	0.1359	5.6080
	SimCLR	0.5629	0.4410	0.6822	0.5823	0.2445	11.3345	0.6217	0.5966	0.8088	0.7618	0.1347	6.3570
	BYOL	0.5819	0.4547	0.6347	0.6644	0.2167	14.0389	0.6291	0.6014	0.8002	0.7803	0.1511	5.9131
	SwAV	0.5993	0.4623	0.6115	0.7142	0.1978	9.9339	0.6346	0.6056	0.8086	0.7795	0.1397	6.7683
	Context	0.6357	0.5081	0.6791	0.6729	0.2504	12.3797	0.6506	0.6244	0.8629	0.7483	0.1475	5.6945
	PCRL	0.6366	0.5086	0.6443	0.7116	0.2388	13.4922	0.6556	0.6273	0.8441	0.7710	0.1463	5.3203
	MG	0.6415	0.5197	0.6497	0.7296	0.2718	9.9594	0.6614	0.6308	0.8519	0.7664	0.1342	5.7091
	SimMIM	0.6422	0.5153	0.6511	0.6857	0.2704	12.4581	0.6665	0.6380	0.8743	0.7482	0.1583	5.9744
	MAE	0.6436	0.5153	0.6444	**0.7440**	0.2655	11.4665	0.6761	0.6469	0.8793	0.7485	0.1586	5.6218
	ADIOS	0.6450	0.5179	0.6676	0.6749	0.2870	12.9934	0.6763	0.6494	0.8883	0.7505	0.1661	5.2881
	ConvMAE	0.6537	0.5333	0.6614	0.7007	0.2802	12.6538	0.6793	0.6516	0.8826	0.7580	0.1353	5.5974
	AttMask	0.6595	0.5322	0.6516	0.7257	0.2831	12.3248	0.6806	0.6500	0.8439	0.7944	0.1478	6.2207
	Ours	**0.6721**	**0.5523**	**0.6837**	0.7421	**0.2923**	**8.8972**	**0.6901**	**0.6678**	**0.8972**	**0.7994**	**0.1664**	**5.1523**
10%	U-Net	0.6493	0.5163	0.6591	0.7227	0.2564	10.9397	0.6777	0.6456	0.8624	0.7643	0.1729	5.3267
	SimCLR	0.6773	0.5663	0.6806	0.7631	0.3026	14.4016	0.6920	0.6595	0.8668	0.7650	0.1922	6.2501
	BYOL	0.6839	0.5765	0.7121	0.7372	0.3193	13.7063	0.7089	0.6784	0.8878	0.7636	0.1722	6.4104
	SwAV	0.6891	0.5773	0.7321	0.7326	0.3180	7.6221	0.7131	0.6791	0.8734	0.7822	0.1832	5.7097
	Context	0.7101	0.6007	0.7342	0.7595	0.3222	7.6234	0.7299	0.6966	0.8601	0.8156	0.1753	5.4416
	PCRL	0.7112	0.6000	0.7126	0.7637	0.3395	9.8741	0.7309	0.6973	0.9035	0.7762	0.2048	6.0055
	MG	0.7116	0.6024	0.6830	0.7856	0.3485	11.951	0.7324	0.6971	0.8721	0.7974	0.1600	5.9994
	SimMIM	0.7124	0.5979	0.7074	0.7767	0.3257	8.8852	0.7327	0.6939	0.8608	0.8081	0.1809	6.1551
	MAE	0.7147	0.5988	0.7104	0.7771	0.3180	7.4786	0.7395	0.7075	0.9169	0.7742	0.2052	5.0874
	ADIOS	0.7160	0.6111	0.7148	0.7655	0.3507	8.5117	0.7402	0.7097	0.9100	0.7883	0.1944	5.3270
	ConvMAE	0.7174	0.6109	0.7245	**0.7919**	0.3492	9.8764	0.7414	0.7086	0.9102	0.7750	0.2056	5.6757
	AttMask	0.7209	0.6090	0.7360	0.7644	0.3412	12.2937	0.7422	0.7100	0.9004	0.7893	0.2143	5.5050
	Ours	**0.7388**	**0.6322**	**0.7389**	0.7868	**0.3517**	**4.8246**	**0.7588**	**0.7205**	**0.9182**	**0.8207**	**0.2188**	**4.8925**
50%	U-Net	0.7222	0.6349	0.7730	0.7279	0.3975	5.8154	0.7407	0.7091	0.8545	**0.8251**	0.2822	4.7658
	Ours	**0.7631**	**0.6713**	**0.7843**	**0.7940**	**0.4442**	**4.9946**	**0.7929**	**0.7665**	**0.9256**	0.8208	**0.2881**	**4.2734**
100%	U-Net	0.7944	0.6941	0.8184	**0.8200**	0.4547	3.9390	0.8316	0.7946	0.8815	**0.9002**	0.2958	3.6435
	Ours	**0.8081**	**0.7122**	**0.8383**	0.8170	**0.4909**	**3.5962**	**0.8565**	**0.8255**	**0.9490**	0.8634	**0.3424**	**3.5857**

labeled data. Firstly, we find that the self-supervised methods are generally better than fully supervised learning from scratch using partial annotations (Fully Supervised). This suggests that, in addition to limited label data, the self-supervised methods also learn useful information from a large amount of unlabeled data. Then, we can see our AHM significantly outperforms the SOTA self-supervised methods in medical image segmentation tasks on both datasets. Specifcally, in the case of 5% annotation, we proposed AHM is 1.91%, 3.56%, 0.67%, and 3.25% higher than the second best result at 5% annotation ratio on the Cardiac dataset for DSC, IoU, PPV, and BIoU, respectively, and HD95 is 1.0367 lower; while at 5% annotation ratio on the TCIA dataset, DSC, IoU, PPV, SEN, and BIoU are 1.40%, 2.48%, 1.00%, 0.62%, and 0.18% higher than the second best result, respectively, and HD95 is 0.1358 lower. In the case of 10% annotation, AHM is 2.48%, 3.45%, 0.39%, and 0.71% higher than the second best result at 10% annotation ratio on the Cardiac dataset for DSC, IoU, PPV, and BIoU, respectively, and HD95 is 2.6540 lower; while at 10% annotation ratio on the TCIA dataset, DSC, IoU, PPV, SEN, and BIoU are 2.24%, 1.48%, 0.14%, 0.62%, and 2.10% higher than the second best result, respectively, and HD95 is 0.1949 lower. The superior performance of AHM can be attributed to the fact that it determines more appropriate and challenging adaptive masks for each image. By reducing the uncertainty of the masked patches, AHM increases the upper limit of conditional mutual information, allowing for the learning of more comprehensive and effective representation information.

3) Analysis of Visualized Segmentation Results: Moreover, all the above findings are also well supported by the visualized results in Figure 4, where our proposed AHM achieves obviously better (i.e., more similar to the ground-truth) segmentation results than all the self-supervised medical image segmentation methods. Specifically, i) the segmentation results of the contrastive learning methods SimCLR, BYOL, and SvAW are highly inaccurate and even over-segmented; ii) Context, SimMIM, MAE, and ConvMAE have better results but unsatisfactory segmentation performance in the edge region (such as the green box); and iii) the proposed AHM segmentation results are closer to the ground truth and retain more details in the foreground region. Thus, these visualization

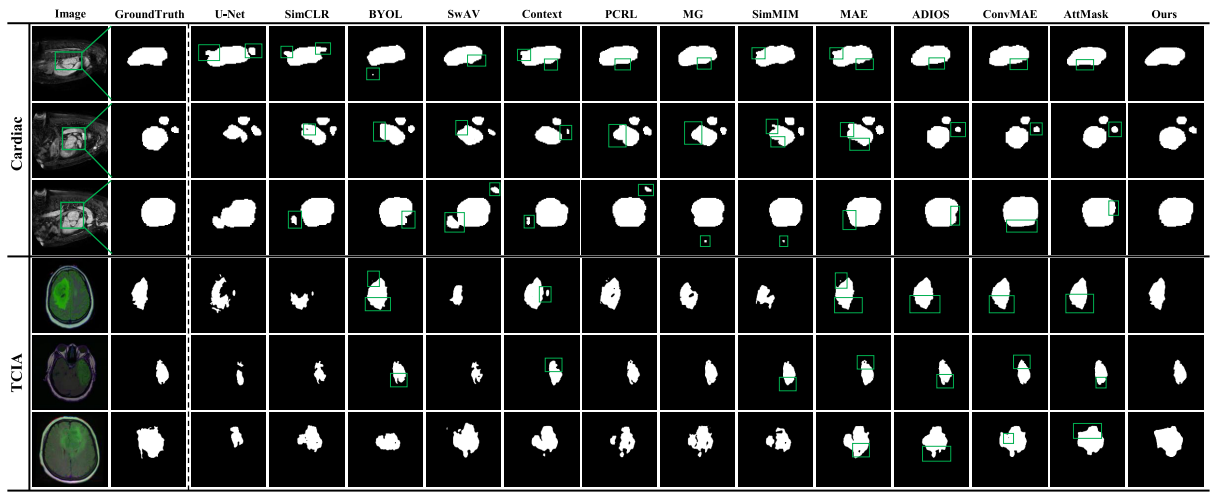


Fig. 4. Examples of visualized segmentation results of our proposed AHM and the baselines on two public datasets.

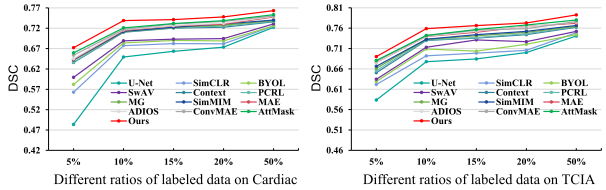


Fig. 5. Segmentation results of the proposed AHM and the other baselines on different ratios labeled data. Our method (red line) outperforms other baseline methods with better generalization performance when using lower proportions of labeled data (i.e., 5%, 10%), and behaviors similarly when using higher ratios of labeled data (i.e., 15%, 20%, 50%) on Cardiac and TCIA dataset.

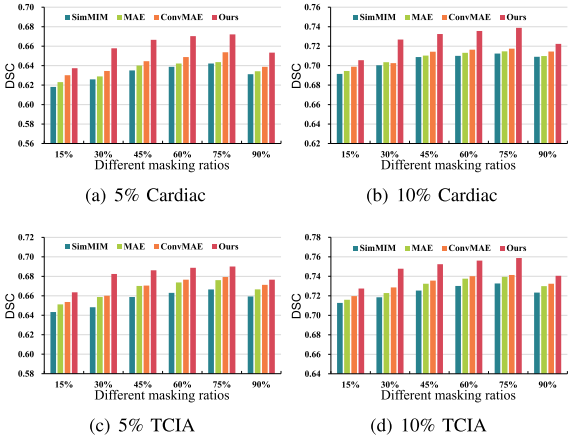


Fig. 6. Comparison of the segmentation results of the proposed AHM and other masking methods on different masking ratio.

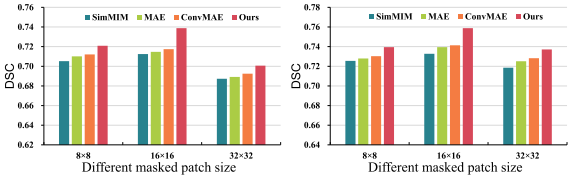
examples demonstrate again that AHM compensates for the shortcomings of existing self-supervised medical image segmentation methods and achieves better performance in medical image segmentation tasks with a small amount of annotation.

E. Ablation Studies

1) Effects on Ratios of Labeled Data: Figure 5 presents the performance gains achieved by our proposed AHM method compared to other baseline methods when fine-tuned on the Cardiac and TCIA datasets using different proportions of labeled data (i.e., 5%, 10%, 15%, 20%, and 50%). We can observe that our method generally outperforms self-supervised baseline methods on all scales. As the proportion of labeled data increases, the performance gap between the methods gradually diminishes. It is worth noting that compared to fine-tuning with high-scale (i.e., 15%, 20%, and 50%) labeled data, the performance gap between other baseline methods and our method is larger in the low-scale (i.e., 5% and 10%) labeled data state. This finding highlights the significant potential of our approach in enhancing model generalization performance and labeling efficiency.

2) Effects on Masking Ratios: we compare the segmentation results of the proposed AHM and other masking methods using different masking ratios on the Cardiac and TCIA datasets, the results are shown in Figure 6. From Figure 6, we can see that our AHM achieves the best performance with a masking rate of 75% on both datasets. Additionally, our method already achieves good performance compared to other methods at a masking rate of 30%, which can be attributed to our method being able to better select patches with more discriminative parts in the images for masking. Additionally, we observe that beyond a masking rate of 30%, the improvement in performance of the downstream segmentation task gradually decreases. This is because at higher masking rates, there are a large number of redundant patches (i.e., background) being sampled, the network just copies features from these patches, resulting in limited improvement in generalization and reduced impact on performance improvement. However, this also indirectly demonstrates that our method is capable of effectively identifying patches with more discriminative parts for masking even at lower masking ratios.

3) Effects on Masking Strategies: In order to validate that more challenging tasks can lead to better performance, we studied various masking strategies in Table V. According to Table V we found that increasing the difficulty of the pretext task does not always result in improved performance, supporting our earlier observation that overly difficult targets are not conducive to learning. Therefore, maintaining a certain



(a) Cardiac (b) TCIA

Fig. 7. Compare the segmentation results of the proposed AHM and other masking methods at different masked patch sizes on Cardiac and TCIA with 10% labeled data.

TABLE V
ABLATION STUDY ON DIFFERENT MASK STRATEGIES. WE STDUY THE EFFECT OF DIFFERENT RANDOM SAMPLING PROBABILITIES p. A SMALLER p INDICATES A MORE DIFFICULT PRETEXT TASK, AND THE STRATEGY WILL RECEIVE FEWER HINTS

p	Difficulty	Cardiac		TCIA	
		5%	10%	5%	10%
0	hard	0.6313	0.7089	0.6678	0.7277
10%		0.6444	0.7136	0.6766	0.7392
20%		0.6621	0.7248	0.6812	0.7476
30%		**0.6721**	**0.7388**	**0.6901**	**0.7588**
40%		0.6702	0.7355	0.6876	0.7513
50%		0.6635	0.7276	0.6802	0.7465
100%	easy	0.6424	0.7135	0.6762	0.7387

level of randomness (i.e., providing some hints) is beneficial for obtaining satisfactory results. Specifically, the best results are obtained at different mask ratios when the random sampling probability p is 30%. When p is 0, it means that the model faces the most challenging problem, where almost all patches with high reconstruction losses are masked, resulting in visible patches being predominantly background. Forcing the model to reconstruct the discriminative parts of the image based solely on this background is meaningless, and its performance continues to decline with different masking ratios. Hence, we can conclude that a certain level of randomness is necessary.

4) Effects on Masked Patch Size: Figure 7 shows the performance of our model and other masking methods are fine-tuned at different masked patch size on the Cardiac and TCIA datasets. From Figure 7, it can be seen that our AHM generally outperforms other methods at different masked patch sizes. Secondly, we can observe that when the masked patch size is 16×16, all methods achieve optimal performance on both datasets. This suggests that a moderate patch size is encouraged in MIM, neither too large nor too small. It is possible that smaller masked patch sizes in MIM may cause the model to learn too many short connections, and too many corresponding hints from smaller visible blocks may limit the improvement in generalization ability. On the other hand, larger masked patch size may make the learning process too difficult.

F. Visualization of Predicted Losses

We provide qualitative results on Cardiac and TCIA validation set in Figure 8. From Figure 8, it can be observed

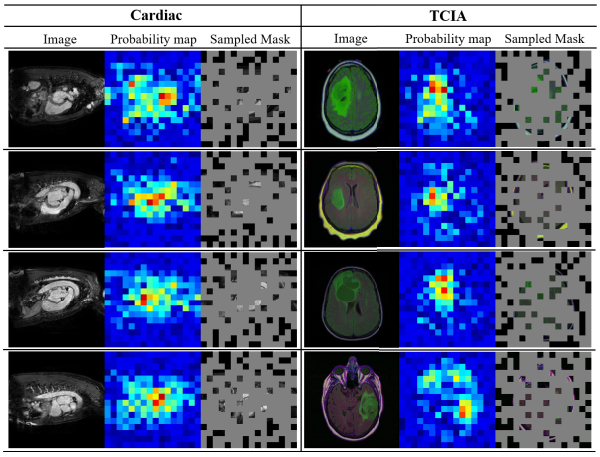
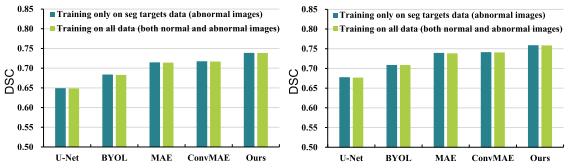


Fig. 8. Visualization on on two public datasets. For each tuple, we show the image (left), predicted reconstruction losses (middle) and masked image generated by sampled prediction loss (right).

that patches with higher predicted reconstruction loss usually are more discriminative (i.e., object or forehead). Moreover, the generated masking strategy based on the predicted reconstruction losses of patches clearly focuses more on regions where organs/lesions exist. This demonstrates that our method effectively increases the difficulty of reconstruction.

G. Additional Experiments

1) Analysis of Not Using Normal Images: In this work, we use only abnormal images (i.e., the images containing segmenting objects) for training and evaluation. The reasons are as follows. First, for evaluation, There is a significant difference in the difficulty of segmentation between images that include the segmentation target (i.e. abnormal images) and images that do not include the segmentation target (i.e. normal images). Abnormal images require the model to accurately depict the edges of the segmentation object, while normal images only require the model to directly output a completely black image. Therefore, we found that different segmentation models have significant differences in performance when segmenting abnormal image; and although false positives may also occur when segmenting normal images, the overall normal image segmentation performance of different models is quite close, and the values are often much greater than when segmenting abnormal images (as outputting a black image is obviously much simpler). As shown in Table II, there are usually many normal images in the medical image dataset, and sometimes even far more than abnormal images (such as TCIA); consequently, we found that if we included normal images for evaluation, it would lead to the segmentation performance of models with poor performance in abnormal image segmentation tasks (segmentation tasks that are truly needed in clinical practice) being erroneously elevated, resulting in BIASED evaluations of the performance of different models in the segmentation tasks (e.g., when there are the same number of normal and abnormal images in the testing set, even a totally useless model that can only output all black images can still obtain a DSC value that is close to 0.5, because although it incorrectly



(a) Cardiac (b) TCIA

Fig. 9. Compare the segmentation results of the proposed AHM with those of baselines on Cardiac and TCIA with 10% labeled data in the following two settings: (i) training using only abnormal images, and (ii) training using both normal and abnormal images.

TABLE VI
RESULTS OF APPLYING THE PROPOSED METHOD AND THE STATE-OF-THE-ART BASELINES ON A PUBLIC NON-CONTRAST CT LUNG SEGMENTATION DATASET, WHERE THE BEST RESULTS ARE IN BOLD, THE SECOND BEST METHODS ARE UNDERLINED

Methods		Lung					
		DSC↑	IoU↑	PPV↑	Sen↑	BIoU↑	HD95↓
10%	U-Net	0.7324	0.6580	0.7767	0.7479	0.4436	25.2898
	SimCLR	0.7528	0.6737	0.7737	0.7749	0.4166	28.6651
	BYOL	0.7750	0.7059	0.9131	0.7286	0.4451	11.3815
	SwAV	0.7899	0.7252	0.8890	0.7579	0.4562	13.2085
	Context	0.7948	0.7348	0.9037	0.7692	0.4929	12.0086
	PCRL	0.8048	0.7377	0.8924	0.7789	0.4832	14.5061
	MG	0.8111	0.7487	0.8803	0.7976	0.4925	13.7226
	SimMIM	0.8223	0.7575	0.8945	0.7947	0.4935	13.5472
	MAE	0.8255	0.7727	0.9032	0.8091	0.5211	10.2076
	ADIOS	0.8316	0.7751	0.8931	0.8166	0.5145	12.3233
	ConvMAE	0.8343	0.7774	0.8911	0.8174	0.5076	12.3039
	AttMask	0.8382	0.7864	0.9183	0.8126	0.5539	8.2652
	Ours	**0.8459**	**0.8029**	**0.9216**	**0.8315**	**0.5875**	**7.5100**
50%	U-Net	0.8363	0.7783	0.9010	0.8176	0.5037	10.8503
	Ours	**0.8617**	**0.8191**	**0.9165**	**0.8475**	**0.5773**	**6.7719**
100%	U-Net	0.8831	0.8374	0.9060	0.8742	0.5934	5.6414
	Ours	**0.9069**	**0.8713**	**0.9251**	**0.9006**	**0.6575**	**3.9668**

segment all abnormal images, its outputs are correct for all normal images, which is obviously biased and unacceptable).

Second, we also find that using normal images in training will bring very little additional information and have only imperceptible impact on the model's segmentation performances. To validate this, we conducted an additional experiment to compare the models' performances when they are trained using only abnormal images with those when using both abnormal and normal images for training. As shown in Figure 9, the results show that on both datasets, the segmentation performances of training only on abnormal images are similar to those of training on both normal and abnormal images. We believe this is because abnormal images also contain lots of normal patches, which have already been enough to provide sufficient information for the model to learn normal features, so using more normal images in training will not bring additional information, and thus bring in little performance enhancement. Considering the significant computational and time costs associated with using normal images in training, and the very limited benefits, we remove them from training in our experiments.

TABLE VII
RESULTS OF APPLYING THE PROPOSED METHOD AND THE RANDOM MASK METHOD ON BRATS2018 DATASET WITH 10% LABELED DATA. † DENOTES THE SEGMENTATION PERFORMANCE WHEN WE TRANSFER THE MASKING STRATEGY TRAINED ON TCIA TO THE SEGMENTATION TASK ON BRATS2018 DATASET, WHERE THE BEST RESULTS ARE IN BOLD, THE SECOND BEST METHODS ARE UNDERLINED

Methods		BraTS2018					
		DSC↑	IoU↑	PPV↑	Sen↑	BIoU↑	HD95↓
10%	MAE	0.7173	0.6180	0.8443	0.7078	0.2680	10.5637
	Ours†	0.7224	0.6248	0.8328	0.7202	0.2634	10.2593
	Ours	**0.7328**	**0.6344**	**0.8491**	**0.7302**	**0.2724**	**9.4023**

2) Analysis of Using Non-Contrast Image: We have additionally introduced a non-contrast CT lung segmentation dataset [36], and validated the effectiveness of our method. The segmentation performance results are presented in Table VI, indicating that our method outperforms other state-of-the-art baselines on this new dataset. This superior performance comes from AHM's ability to select adaptive masks tailored to the characteristics of medical images. Particularly, the performance of AHM with 10% annotations surpasses that of the fully supervised method with 50% annotations. The segmentation visualizations in Figure 10 further illustrate that even when using non-contrast images, our method is capable of determining more suitable and challenging adaptive masks for each image. This helps the model learn more comprehensive and effective generalized information, thereby enhancing performance in downstream tasks.

3) Effect of the Quality of Reconstruction Network: We additionally study the impact of different qualities of reconstruction networks on the performances of our method. We selected reconstruction networks trained at various epochs, each with different reconstruction losses to aid the training of our AHM. As depicted in Figure 11, the values in loss indicate the quality of the reconstruction network, and the values in DSC indicate the performances of our model. Consequently, it is very obvious that the quality of reconstruction network have strong and direct correlation with the performances of our model: when the loss is high (i.e., the quality of reconstruction is low), the performance of our model is bad, when the loss continue to rise and keep steady at around 0.004 (resp., 0.0043) on Cardiac (resp., TCIA), the DSC of our model also rise to around 0.7388 (resp., 0.7588). Therefore, for optimal training of the A3C masking network, we should first train the reconstruction network to converge, i.e., with the value of loss at around 0.004 (resp., 0.0043) on Cardiac (resp., TCIA).

4) Analysis of Transfer of Masking Strategy: We further study whether the masking strategy learned by AHM can be transferred to other datasets with the same tasks. Here, we addtionally introduce a new brain tumor MRI segmentation dataset, BraTS2018 [37], [38], [39], due to the fact that both BraTS2018 and TCIA datasets involve the segmentation of brain tumors, we choose BraTS2018 as our transfer dataset, and transfer the masking strategy trained on TCIA to BraTS2018, and the experimental results are presented in

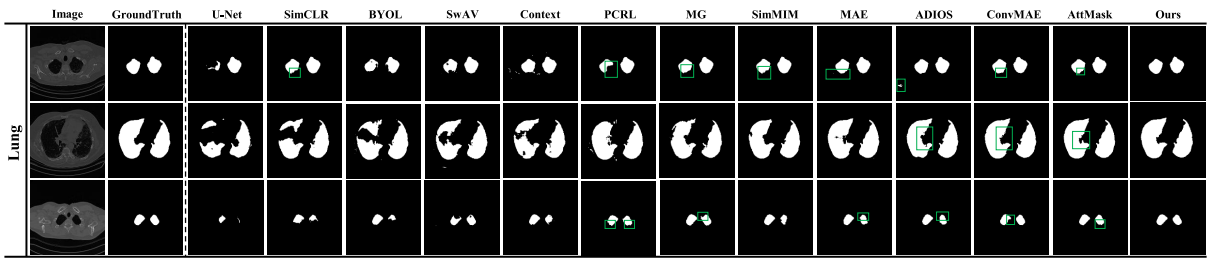
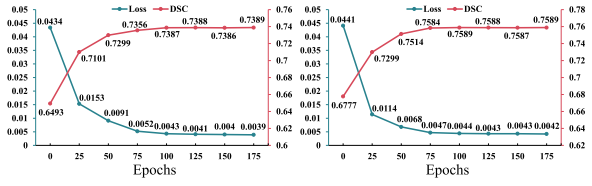
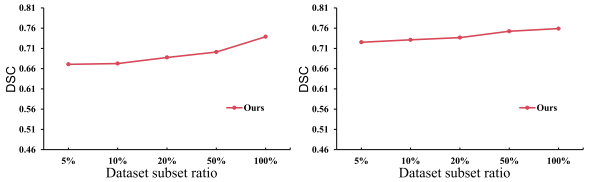


Fig. 10. Examples of visualized segmentation results of our proposed AHM and the baselines on a public non-contrast CT lung segmentation dataset.



(a) Cardiac (b) TCIA

Fig. 11. Compare the segmentation results of training our proposed AHM with reconstruction networks trained at different epochs on Cardiac and TCIA with 10% labeled data.



(a) Cardiac (b) TCIA

Fig. 12. Compare the segmentation results of our proposed AHM trained on subsets with different ratios of the training set, and evaluate on Cardiac and TCIA with 10% labeled data.

Table VII, where † indicates the segmentation performance when we transfer the masking strategy trained on TCIA to the segmentation task on BraTS2018. From Table VII, it can be observed that our AHM, using strategies transferred from TCIA to BraTS2018, achieved better segmentation performance on most metrics compared to the random masking method trained on BraTS2018. Although the segmentation performance of the transferred AHM is not as high as that of the AHM trained on BraTS2018, it still demonstrates competitive results. This indicates that the masking strategy learned by AHM has a certain degree of generalization, allowing it to transfer learned strategies from one dataset to another similar dataset for the same task.

5) Analysis of Training on Small Subsets: In order to study whether AHM can be effectively trained on smaller training subsets, we attempted to train AHM on training subsets with different ratios (5%, 10%, 20%, 50%, and 100%) of the training set. As shown in Figure 12, even when trained on very small subsets, AHM did not experience a significant drop in performance and was able to converge well. Additionally, as the dataset size increased, there was a noticeable improvement in segmentation performance. This suggests that AHM does not have strict requirements on the size of the training dataset and can learn effective representations even when trained on smaller datasets. This flexibility is advantageous for the extension of AHM to different datasets and tasks.

TABLE VIII
CLASSIFICATION RESULTS OF APPLYING THE PROPOSED METHOD AND REPRESENTATIVE CONTRASTIVE LEARNING AND MIM BASELINES ON MESSIDOR-2 DATASET

Methods		Messidor-2			
		Acc↑	Pre↑	Rec↑	F1↑
10%	ResNet-50	0.6914	0.6439	0.5822	0.6115
	BYOL	0.7314	0.7031	0.6164	0.6569
	MAE	0.7686	0.6901	0.7603	0.7475
	ConvMAE	0.7743	0.7081	0.7808	0.7427
	Ours	**0.7971**	**0.7419**	**0.7877**	**0.7641**
50%	ResNet-50	0.7914	0.7355	0.7808	0.7575
100%	ResNet-50	0.8343	0.8188	0.7740	0.7958

6) Analysis of Detection and Classification Tasks: We validated the effectiveness of AHM on downstream classification and detection tasks. Specifically, We validated our proposed AHM alongside representative contrastive learning and MIM baselines on the public classification dataset Messidor-2 [40], [41], [42] for diabetic retinopathy, utilizing widely adopted metrics such as accuracy (Acc), precision (Pre), recall (Rec), and F1 score (F1) for evaluation. The results for the classification task (presence of disease or absence of disease) are presented in Table VIII, From Table VIII, we can see that due to the ability of AHM to learn key features more precisely, our method is still better than other baselines in the classification task. Similar to the segmentation results, the performance of AHM with 10% annotations still surpasses that of fully supervised methods with 50% annotations.

For the detection task, we transformed the segmentation labels of the publicly available TCIA dataset into detection labels to evaluate the performance of AHM and representative contrastive learning and MIM baselines in the detection task. We employed widely used metrics such as Average Precision (AP) and Sensitivity (Sen) for evaluation. The detection results are presented in Table IX, indicating that our method outperforms other baselines in the detection task as well. This underscores the capability of our proposed AHM to learn more comprehensive and effective generalization information, thereby enhancing performance in downstream tasks.

TABLE IX
DETECTION RESULTS OF APPLYING THE PROPOSED METHOD AND REPRESENTATIVE CONTRASTIVE LEARNING AND MIM BASELINES ON TCIA DATASET

Methods		TCIA			
		AP↑		Sen↑	
		@50	@75	0.5	4
10%	Faster-RCNN	0.6802	0.2838	0.7159	0.7509
	BYOL	0.7068	0.3462	0.7228	0.7840
	MAE	0.7411	0.3557	0.7581	0.8064
	ConvMAE	0.7594	0.3806	0.7842	0.8229
	Ours	**0.7859**	**0.4139**	**0.8053**	**0.8431**
50%	Faster-RCNN	0.7908	0.4714	0.8269	0.8526
100%	Faster-RCNN	0.8142	0.6726	0.8485	0.8863

VI. DISCUSSION

We now briefly summarize the social impact of our approach, as well as the limitations of our work and future works.

A. Social Impact of Proposed Approach

Medical image classification, detection, and segmentation have a significant impact on subsequent medical diagnosis and clinical tasks. However, acquiring annotated medical data in practice is a time-consuming and labor-intensive task, greatly limiting the efficiency and performance of computer aided diagnosis (CAD) in clinical practice. Particularly, obtaining pixel-level masks for medical segmentation tasks is even more costly. Therefore, many approaches have been introduced that leverage a large amount of unlabeled medical data through self-supervised methods to improve the performance of downstream tasks. However, these methods overlook the differences between natural images and medical images. Although existing self-supervised methods can achieve certain improvements when directly applied to medical images, the gains are limited.

To address this, we propose AHM, a self-supervised masking strategy based on deep reinforcement learning. AHM is designed to adapt to the characteristics of medical images by selecting appropriate and challenging masks for each medical image. This enhances the model's learning ability for key features and, in turn, improves the performance of downstream tasks. Additionally, our self-supervised masking strategy can be seen as a framework that can easily be integrated into other methods. Its network structure can be replaced with other models for feature extraction, and it can also add any improved modules.

Therefore, in addition to the technical contributions, our work also brings significant social benefits in related research and clinical fields. For example, by leveraging unlabeled data to reduce annotation costs, especially in the field of medical image segmentation, our approach has several advantages. Firstly, it accelerates the application process of intelligent computer-aided diagnosis systems. Secondly, it alleviates the problem of convergence difficulties when training with small annotated medical image datasets, significantly reducing the workload for doctors and saving time and labor costs. By addressing these challenges, our work has the potential to enhance the efficiency and effectiveness of medical image analysis, ultimately benefiting both medical professionals and patients.

B. Resolving Overfitting Problem

In medical image datasets, normal images can significantly outnumber abnormal images, which sometimes may results in overfitting problem. However, in our work, as discussed in Subsection V-G, we find that using normal images in training will bring very little additional information and have only imperceptible impact on the model's segmentation performances, so we remove them from training in our experiments, which thus also avoid the potential overfitting problem due to too many normal images.

Actually, the potential overfitting problem in our work comes from the fact that the masking strategy may mask too many hard patches, making the model only learn from easy patches and overfit to normal cases. To address this, we incorporated a randomness strategy in our work. During masking, a certain number of hard patches are randomly selected to be retained and not allowed to be masked. This ensures that the resulting image after masking retains enough hard patches (i.e., some degree of hints) to provide sufficient information for subsequent image reconstruction. The experimental results in Table V greatly support this argument: we compared the model's performance with and without the randomness strategy, and the results indicate that the model performs significantly better with the randomness strategy ($p = 30\%$) than without it ($p = 0$).

C. Limitations and Future Work

Although our experiments primarily focused on 2D medical images, we believe that the proposed AHM can also be used for 3D image patches as well. Given the wealth of information in 3D images, generating a mask strategy in a single pass may not be the most optimal. Hence, a potential way of extending AHM to 3D tasks is to transform the mask strategy into a multi-iteration form, allowing the network to iteratively optimize from coarse to fine to learn the best mask strategy. However, the time and computation cost for 3D extension is very expensive, which requires high-power GPU with large GPU RAM, e.g., A100 or H800 GPUs, and will takes quite long time for training, so we leave the extension of self-supervised learning in 3D medical images as a potential future work. Furthermore, mask ratios have traditionally been manually set hyperparameters based on human experience. In some medical datasets, there can be significant variations in the sizes of organs or tumor regions. Larger organs or tumor areas might require higher mask ratios, while smaller ones may suffice with lower mask ratios. Therefore, if we can use deep reinforcement learning frameworks to dynamically select an appropriate mask ratio for each image, it might help better capture key features in datasets with such significant variations. This could also potentially reduce training costs to some extent. Hence, this is an interesting research direction we plan to explore in the future.

VII. Conclusion

In this work, we first identified the limitations of existing self-supervised methods in medical imaging due to higher redundancy and smaller regions of key organs/lesions in medical images compared to natural images. To address this issue, we propose AHM, which introduces a deep reinforcement learning model to predict reconstruction loss and select appropriate and challenging masks for each image. This effectively helps the reconstruction network learn more fine-grained image representations in regions with organs/lesions, thereby improving the performance of downstream segmentation models. We conducted extensive experiments on the Cardiac and TCIA datasets, and the results demonstrate that our method outperforms current state-of-the-art self-supervised methods. Ablation studies also confirm the effectiveness and superiority of the proposed AHM.

References

[1] S. Shurrab and R. Duwairi, "Self-supervised learning methods and applications in medical imaging analysis: A survey," 2021, *arXiv:2109.08685*.

[2] H. Bao, L. Dong, S. Piao, and F. Wei, "BEiT: BERT pre-training of image transformers," 2021, *arXiv:2106.08254*.

[3] B. Cheng, R. Girshick, P. Dollár, A. C. Berg, and A. Kirillov, "Boundary IoU: Improving object-centric image segmentation evaluation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15334–15342.

[4] T. Brown et al., "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 1877–1901.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, vol. 1, 2019, p. 2.

[6] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 16000–16009.

[7] Z. Xie et al., "SimMIM: A simple framework for masked image modeling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 9653–9663.

[8] P. Gao, T. Ma, H. Li, Z. Lin, J. Dai, and Y. Qiao, "ConvMAE: Masked convolution meets masked autoencoders," 2022, *arXiv:2205.03892*.

[9] H. Wang, K. Song, J. Fan, Y. Wang, J. Xie, and Z. Zhang, "Hard patches mining for masked image modeling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 10375–10385.

[10] V. Mnih et al., "Asynchronous methods for deep reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1928–1937.

[11] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[12] J.-B. Grill et al., "Bootstrap your own latent—A new approach to self-supervised learning," in *Proc. NeurIPS*, vol. 33, 2020, pp. 21271–21284.

[13] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 9912–9924.

[14] H.-Y. Zhou, C. Lu, S. Yang, X. Han, and Y. Yu, "Preservational learning improves self-supervised medical image models by reconstructing diverse contexts," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3499–3509.

[15] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2536–2544.

[16] Z. Zhou et al., "Models genesis: Generic autodidactic models for 3D medical image analysis," in *Proc. Med. Image Comput. Comput. Assist. Intervent.*, 2019, pp. 384–393.

[17] Y. Shi, N. Siddharth, P. Torr, and A. R. Kosiorek, "Adversarial masking for self-supervised learning," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 20026–20040.

[18] I. Kakogeorgiou et al., "What to hide from your students: Attention-guided masked image modeling," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 300–318.

[19] Y. Song et al., "Mega-reward: Achieving human-level play without extrinsic rewards," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 4, pp. 5826–5833.

[20] D. Yuan, Y. Liu, Z. Xu, Y. Zhan, J. Chen, and T. Lukasiewicz, "Painless and accurate medical image analysis using deep reinforcement learning with task-oriented homogenized automatic pre-processing," *Comput. Biol. Med.*, vol. 153, Feb. 2023, Art. no. 106487.

[21] G. Xu, S. Wang, T. Lukasiewicz, and Z. Xu, "Adaptive-masking policy with deep reinforcement learning for self-supervised medical image segmentation," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2023, pp. 2285–2290.

[22] L. Wang, K. Lekadir, S.-L. Lee, R. Merrifield, and G.-Z. Yang, "A general framework for context-specific image segmentation using reinforcement learning," *IEEE Trans. Med. Imag.*, vol. 32, no. 5, pp. 943–956, May 2013.

[23] X. Liao et al., "Iteratively-refined interactive 3D medical image segmentation with multi-agent reinforcement learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9394–9402.

[24] Z. Tian, X. Si, Y. Zheng, Z. Chen, and X. Li, "Multi-step medical image segmentation based on reinforcement learning," *J. Ambient Intell. Humanized Comput.*, vol. 13, pp. 5011–5022, Mar. 2020.

[25] Y. Man, Y. Huang, J. Feng, X. Li, and F. Wu, "Deep Q learning driven CT pancreas segmentation with geometry-aware U-Net," *IEEE Trans. Med. Imag.*, vol. 38, no. 8, pp. 1971–1980, Aug. 2019.

[26] T. Qin, Z. Wang, K. He, Y. Shi, Y. Gao, and D. Shen, "Automatic data augmentation via deep reinforcement learning for effective kidney tumor segmentation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 1419–1423.

[27] Z. Li and Y. Xia, "Deep reinforcement learning for weakly-supervised lymph node segmentation in CT images," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 3, pp. 774–783, Mar. 2021.

[28] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[29] R. Furuta, N. Inoue, and T. Yamasaki, "PixelRL: Fully convolutional network with reinforcement learning for image processing," *IEEE Trans. Multimedia*, vol. 22, no. 7, pp. 1704–1719, Jul. 2020.

[30] A. L. Simpson et al., "A large annotated medical image dataset for the development and evaluation of segmentation algorithms," 2019, *arXiv:1902.09063*.

[31] M. Buda, A. Saha, and M. A. Mazurowski, "Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm," *Comput. Biol. Med.*, vol. 109, pp. 218–225, Jun. 2019.

[32] Q. Yu, Y. Xia, L. Xie, E. K. Fishman, and A. L. Yuille, "Thickened 2D networks for efficient 3D medical image segmentation," 2019, *arXiv:1904.01150*.

[33] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241.

[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, 2014, pp. 1–15.

[35] D. Karimi and S. E. Salcudean, "Reducing the Hausdorff distance in medical image segmentation with convolutional neural networks," *IEEE Trans. Med. Imag.*, vol. 39, no. 2, pp. 499–513, Feb. 2020.

[36] (2020). *Lung Segmentation Dataset*. [Online]. Available: https://www.kaggle.com/sandorkonya/ct-lung-heart-trachea-segmentation

[37] B. H. Menze et al., "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE Trans. Med. Imag.*, vol. 34, no. 10, pp. 1993–2024, Oct. 2015.

[38] S. Bakas et al., "Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features," *Sci. Data*, vol. 4, no. 1, 2017, Art. no. 170117.

[39] S. Bakas et al., "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge," 2018, *arXiv:1811.02629*.

[40] E. Decencière et al., "Feedback on a publicly distributed image database: The Messidor database," *Image Anal. Stereol.*, vol. 33, no. 3, pp. 231–234, 2014.

[41] J. Krause et al., "Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy," *Ophthalmology*, vol. 125, no. 8, pp. 1264–1272, 2018.

[42] V. Gulshan et al., "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *JAMA*, vol. 316, no. 22, pp. 2402–2410, 2016.