MuRCL: Multi-Instance Reinforcement Contrastive Learning for Whole Slide Image Classification

Zhonghang Zhu, Lequan Yu, *Member, IEEE*, Wei Wu, Rongshan Yu, Defu Zhang, and Liansheng Wang, *Member, IEEE*

***Abstract*— Multi-instance learning (MIL) is widely adopted for automatic whole slide image (WSI) analysis and it usually consists of two stages, i.e., instance feature extraction and feature aggregation. However, due to the "weak supervision" of slide-level labels, the feature aggregation stage would suffer from severe over-fitting in training an effective MIL model. In this case, mining more information from limited slide-level data is pivotal to WSI analysis. Different from previous works on improving instance feature extraction, this paper investigates how to exploit the latent relationship of different instances (patches) to combat overfitting in MIL for more generalizable WSI classification. In particular, we propose a novel Multi-instance Reinforcement Contrastive Learning framework (MuRCL) to deeply mine the inherent semantic relationships of different patches to advance WSI classification. Specifically, the proposed framework is first trained in a self-supervised manner and then finetuned with WSI slide-level labels. We formulate the first stage as a contrastive learning (CL) process, where positive/negative discriminative feature sets are constructed from the same patch-level feature bags of WSIs. To facilitate the CL training, we design a novel reinforcement learning-based agent to progressively update the selection of discriminative feature sets according to an online reward for slide-level feature aggregation. Then, we further update the model with labeled WSI data to regularize the learned features for the final WSI classification. Experimental results on three public WSI classification datasets (Camelyon16, TCGA-Lung and TCGA-Kidney) demonstrate that the proposed MuRCL outperforms state-of-the-art MIL models. In addition, MuRCL can achieve comparable performance to other state-of-the-art MIL models on TCGA-Esca dataset.**

Manuscript received 27 September 2022; revised 23 November 2022; accepted 26 November 2022. Date of publication 7 December 2022; date of current version 2 May 2023. This work was supported by the National Key Research and Development Program of China under Grant 2019YFE0113900. *(Zhonghang Zhu and Lequan Yu contributed equally to this work.) (Corresponding author: Liansheng Wang.)*

Zhonghang Zhu is with the School of Informatics, National Institute for Data Science in Health and Medicine, Xiamen University, Xiamen 361005, China (e-mail: zzhonghang@stu.xmu.edu.cn).

Lequan Yu is with the Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong, SAR, China (e-mail: lqyu@hku.hk).

Wei Wu, Rongshan Yu, Defu Zhang, and Liansheng Wang are with the School of Informatics, Xiamen University, Xiamen 361005, China (e-mail: weiwu@stu.xmu.edu.cn; rsyu@xmu.edu.cn; dfzhang@xmu.edu.cn; lswang@xmu.edu.cn).

Digital Object Identifier 10.1109/TMI.2022.3227066

***Index Terms*— Whole slide image analysis, multi-instance learning, contrastive learning, reinforcement learning.**

I. Introduction

WHOLE slide images (WSIs) are gold standard of cancer diagnosis [1], [2]. WSI analysis is also critical to the study of disease onset and progression, and the development of target therapies [3]. While automated WSI analysis is a long-standing challenging problem in medical image analysis due to the high resolutions of WSIs (*e.g.*, the typical size is 40,000 × 40,000). There were some recent successes in computer vision and medical image analysis communities for WSI analysis [4], [5], [6]. Among them, multi-instance learning (MIL) is widely adopted when only slide-level labels are available [7], [8], where the WSI is considered as a bag that contains many instances of image patches. In the typical MIL-based WSI analysis pipeline, patch-level features are extracted, aggregated, and classified to generate the slide-level prediction.

Recent MIL-based approaches have greatly benefited from deep neural networks for feature extraction and aggregation with slide-level supervision [8], [9]. However, several challenges still exist in developing robust models for accurate WSI analysis. First, as mentioned above, most of the existing MIL methods rely on weakly-supervised slide-level labels to train the whole framework. Due to the limited number of whole slide images, these models would easily suffer from overfitting and are unable to learn rich representations due to the weak supervisory signal [10], [11]. Second, as the end-to-end MIL training process is prohibitively expensive for large feature bags like WSIs, current models only sample a few high score patch features [4], [10], [11], [12] for slide-level prediction to reduce computation cost. However, these models would lead to sub-optimal solutions for WSI classification, as this learning scheme fails to consider semantic relationship among patches for discriminative feature selection. How to select discriminative patch features to train the MIL feature aggregator is pivotal to accurate WSI classification, especially under the limited training data scenario.

To tackle the above challenges, we aim to investigate an effective learning paradigm to explore the valuable semantic relationship of different patches in WSIs for generalizable slide-level WSI analysis, where discriminative patch features can also be automatically selected to enhance the performance

1558-254X © 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: SICHUAN UNIVERSITY. Downloaded on February 23,2024 at 07:42:13 UTC from IEEE Xplore. Restrictions apply.

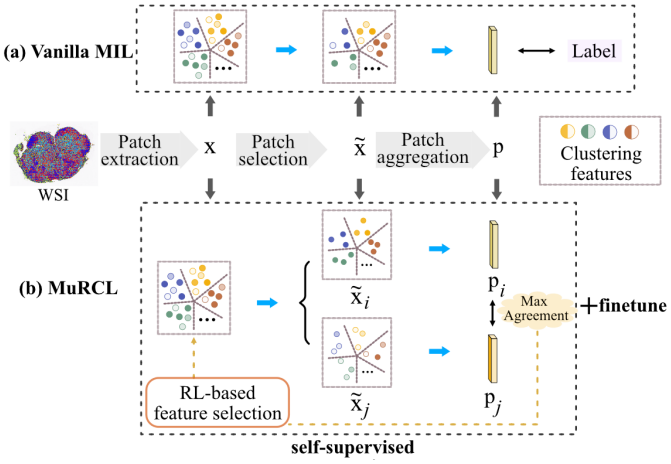

Fig. 1. Comparison between the vanilla MIL method and our MuRCL. **(a)**: A common MIL with patch extraction, patch selection, and patch aggregation is trained with image-level label. **(b)**: Our MuRCL is trained by maximizing the agreement between two discriminative feature sets of the same WSI to exploit the inherent relationship of different patches, followed by a fine-tune process for final WSI prediction.

of the framework. As shown in Fig. 1, a vanilla MIL process (Fig. 1(a)) can be viewed as a three-component paradigms: patch extraction, patch selection and patch aggregation. The WSI feature bag x is extracted from the given WSI and then a subset patch feature $\tilde{x}$ is selected from x for feature aggregation to generate the slide-level prediction p. To mine the inherent semantic relationship of WSI patches, we propose a novel multi-instance contrastive learning framework (Fig. 1(b)), where two discriminative feature sets are exploited from the same WSI and then the framework is trained to maximize the agreement between these two sets in the self-supervised phase, followed by a finetune process for final WSI prediction. Note that our approach is different from the previous patch-level self-supervised method [12], which is dedicated to training a feature extraction network in a self-supervised manner, whereas we focus on how to self-supervisedly exploit the relationship of different patches to advance slide-level representation on top of the extracted patch features.

In this paper, we propose a novel **Mu**lti-instance **R**einforcement **C**ontrastive **L**earning (MuRCL) framework to advance WSI classification. Our proposed framework is first trained in a self-supervised manner to exploit the inherent semantic relationships of different patches and then we fine-tune it for final prediction. We formulate the self-supervised phase as a CL process, where positive/negative discriminative feature sets are constructed from the patch-level feature bags of WSIs. To select informative patch features from WSIs, we further design a novel reinforcement learning (RL) agent to guide the discriminative feature set construction. Particularly, given a WSI patch-level feature bag, we maintain an agent triggered by different initial states to separately construct two discriminative feature sets from it to form the positive pair in CL training, where the agent will progressively update the selection of the discriminative feature set over training iterations according to an online reward. We employ the cosine similarity between input positive feature sets as the reward, which aims to guide the agent to retrieve distinct features and thus make selected feature sets more informative for CL. We further fine-tune the framework with WSI labels for the final WSI classification. It is worth noting that we didn't merely combine different existing modules into a framework. Instead, we proposed a novel and comprehensive solution by seamlessly adapting optimization strategy in RL and CL to an important problem in WSI analysis, *i.e.*, self-supervised MIL aggregation. Also, how to design different informative views of the same sample to enable set-based CL is non-trivial and we for the first time to integrate RL into our set-based CL framework by training the RL module to propose distinct views to maximize the designed reward, *i.e.*, the CL objective.

Our main contributions can be summarized as follows:

- We present a novel multi-instance contrastive learning framework to advance WSI classification by mining the inherent semantic relationships of different patches.
- We propose a novel set-based positive pair construction solution to enable the self-supervised multi-instance contrastive learning.
- We propose a novel RL-based strategy to select discriminative sets from WSI. An agent is designed to dynamically refine the set selection for self-supervised MIL contrastive learning, which is trained by reinforcement learning.
- We have validated our framework on four benchmark WSI datasets. Our method outperforms previous SOTA methods on Camelyon16, TCGA-Lung and TCGA-Kidney, and achieves comparable performance to SOTA methods on TCGA-Esca. Code is at available https://github.com/wwu98934/MuRCL.

II. Related Works

A. MIL for WSI Classification

MIL is widely used in various applications where multiple instances are observed and only one general category label is given, including WSI analysis. It is usually to divide a WSI into small patches that could be considered as a bag with a single label [13], which can be implemented using MIL approaches. Recently, MIL has been successfully applied to WSI analysis [3], [13], [14], [15]. There are two main streams of MIL approaches [16]: the instance-level approach [4] and the embedding-level approach [17]. In the first category, a model is trained by assigning each instance a pseudo-label based on the bag-level label and selecting the top-k instances for aggregation. For the second category, patches of the slide are firstly mapped to a feature space, and then are aggregated by an operator. Most of the current MIL approaches are based on the second one and have achieved good performance for WSI analysis. However, these methods easily suffer from overfitting when training data is limited. We thus propose a self-supervised MIL-based approach to boost WSI analysis.

B. Self-Supervised Learning

Self-supervised learning (SSL) has achieved considerable success in natural and medical image analysis tasks [18], [19], [20]. Existing SSL approaches construct different pretext tasks to exploit supervisory information from the input data. They can be grouped into three categories [18], [21]: 1) context-based, 2) generative-based, and 3) contrastive-based methods. For the first category, the design of pretext tasks is generally

Authorized licensed use limited to: SICHUAN UNIVERSITY. Downloaded on February 23,2024 at 07:42:13 UTC from IEEE Xplore. Restrictions apply.

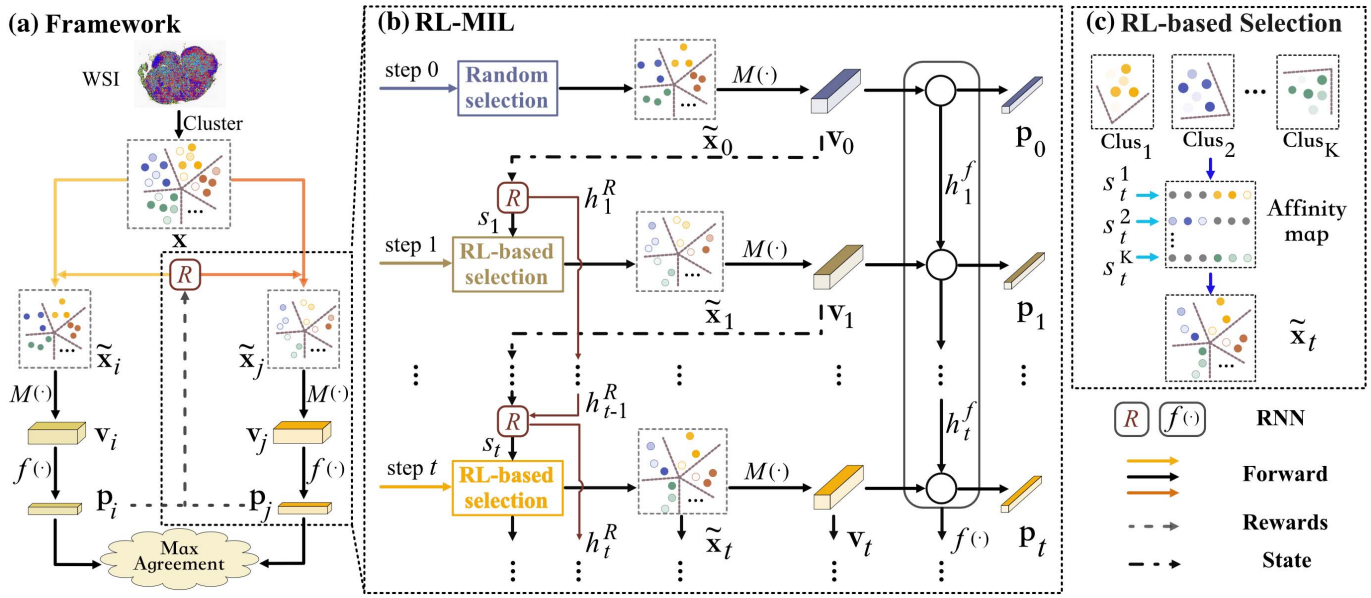

Fig. 2. Illustrations of the self-supervised learning phase of the proposed MuRCL. **(a)** MuRCL: given the input feature bag (WSI-Fbag) x of a WSI, the outputs of two RL-MIL branches are made of positive pairs for contrastive loss training. **(b)** RL-MIL: a reward-guided agent R is employed to select the discriminative feature set (WSI-Fset) $\tilde{x}$ from x. $\tilde{x}$ is randomly selected at the first step and then is determined by the agent R in following steps. $\tilde{x}$ is further sent to the MIL aggregator $M(\cdot)$ to generate the feature embedding v. Outputs p after the projection head $f(\cdot)$ will be used to compute the contrastive loss. **(c)** RL-based Selection: the proposed s_t is applied to pick out indexed features from each cluster of the affinity map with s_t^k, $k \in [1, 2, \ldots, K]$.

based on domain-specific knowledge according to the data, like segmentation task [22], [23], [24], [25] and classification task [19], [26]. For the second category, models are regularized by minimizing the reconstruction loss in the pixel space [27], [28]. The third category refers to a class of newly proposed methods, which learns to enforce similarities in the latent space between positive/negative pairs [29], [30]. Notable techniques have been proposed to extend the contrastive learning to medical image analysis [11], [31], while to our best knowledge, there are no works on multi-instance contrastive learning to learn how to conduct feature aggregation.

C. Reinforcement Learning on Visual Tasks

Deep learning is widely been investigated in the domain of RL. By introducing deep neural networks to build the policy function, the RL agent has a powerful ability to learn the dynamics of the environment with the interactions [32]. Recently, some RL-based approaches have been proposed for visual image processing [33], [34], which has achieved prominent performance. In addition, RL-based methods have been proposed to automatically find an effective augmentation for classification task [35] or generalization in RL [36], [37], [38]. However, few works employ reinforcement learning for promoting contrastive learning. To the best of our knowledge, this is the first time to adopt reinforcement learning to promote CL-based representation learning. Significantly, different from previous work that combines CL with RL *i.e.*, CURL [38], our framework utilizes RL to dynamically construct positive/negative pairs to facilitate set-based CL learning, while CURL takes CL as a self-supervised auxiliary task to promote the feature representation to improve the performance of RL tasks.

III. Methodology

Fig. 2(a) overviews the framework of the proposed MuRCL, where we aim to select two separate discriminative sets from an input feature bag of a WSI by an agent to construct positive/negative pairs for contrastive learning. Then a MIL aggregator $M(\cdot)$ and a projection head $f(\cdot)$ are trained to maximize the agreement of the positive discriminative sets using a contrastive loss. Each branch of the MuRCL is termed as RL-MIL. In Fig. 2(b), we illustrate the sequential decision process of the RL-MIL. With the input feature bag, the RL-MIL iteratively produces a sequence of discriminative sets and outputs a feature vector sequence with the MIL aggregator and projection head. Specifically, at each step, a group of feature indexes is determined by a reward-guided agent (*i.e.*, the discriminative set proposal network). Then, the discriminative set for the next step is constructed by compounding the indexed features of each cluster of input feature bag as shown in Fig. 2(c). Finally, the self-supervised pretrained model will be finetuned with WSI labels to regularize the learned representations for final prediction.

A. Multiple Instance Contrastive Learning

Our multiple instance contrastive learning framework takes the WSI feature bag (WSI-Fbag) as input, where the feature bag consists of patch-level embeddings extracted by a ResNet18 [39] pre-trained on ImageNet. One key component of CL is to construct logical positive/negative pairs (*i.e.*, semantically similar/dissimilar examples) for training. Different from the previous image augmentation-based strategy, we propose to sample different WSI discriminative sets (WSI-Fset for short) from each WSI-Fbag to construct set-based positive/negative pairs for CL training.

Authorized licensed use limited to: SICHUAN UNIVERSITY. Downloaded on February 23,2024 at 07:42:13 UTC from IEEE Xplore. Restrictions apply.

Each WSI-Fset is a combination of sub-sets from multiple feature clusters of the WSI-Fbag. Specifically, given a WSI-Fbag x. We first use a clustering algorithm (*e.g.*, K-means [40]) to split the WSI-Fbag x into K feature clusters C_k (k∈[1,2,…,K]). By sampling a sub-cluster from k-th cluster, the compound WSI-Fset $\tilde{x}$ can be formed as the concatenation of these sub-clusters. We use the same sampling ratio for each cluster, so that the constructed WSI-Fset has constant number of instance embeddings. After getting the WSI-Fset $\tilde{x}$, we feed it into a MIL aggregator $M(\cdot)$ following a projection head $f(\cdot)$ to produce the WSI-level feature embedding p. It is worth noting that with different sampling strategies, we can obtain different WSI-Fsets and the corresponding WSI-level embeddings from the same WSI-Fbag. These different WSI-Fsets can be regarded as *different views* of the same WSI and thus can be used to form positive pairs in CL.

Suppose that $\{\tilde{x}_n\}_{n=1}^{N}$ is a group of N WSI-Fsets, where $\tilde{x}_i$ and $\tilde{x}_j$ are sampled from the same WSI-Fbag and the others are from different WSI-Fbags. Then, the CL loss can be calculated as [29]:

$$L_{i,j} = -\log \frac{\exp(sim(p_i, p_j)/\tau)}{\sum_{n=1}^{N} \mathbb{1}(n \neq i) \cdot \exp(sim(p_i, p_n)/\tau)}, \quad (1)$$

The τ represents the temperature parameter and the $sim(\cdot,\cdot)$ denotes the cosine similarity of two vectors. while the $\mathbb{1}(n \neq i) \in \{0, 1\}$ is an indicator function evaluating to 1 if $n \neq i$. In this work, we employ the NT-Xent loss [29] as the objective function to maximize the similarity between the positive pair and minimize the similarity between the negative pairs. In this case, the MIL aggregator will be encouraged to learn the aggregation knowledge for accurate prediction.

B. RL-Driven Discriminative Set Construction

As aforementioned, how to construct the WSI discriminative set (WSI-Fset) is important in MuRCL. We thus propose a novel reinforcement learning-driven strategy, termed RL-MIL, to construct the discriminative WSI-Fset from the WSI-Fbag, where a WSI-Fset proposal agent R (*i.e.*, a recurrent neural network) is trained by reinforcement learning. As shown in Fig. 2(b), the WSI-Fset construction can be formulated as a sequential decision process. At each step, the MIL aggregator $M(\cdot)$ and the projection head $f(\cdot)$ take a WSI-Fset as input and produce the corresponding semantic prediction p. Meanwhile, the agent R will also generate another WSI-Fset proposal s for the next step by taking the feature vector v as the input.

Specifically, given the input WSI-Fbag x, the RL-MIL module iteratively processes a sequence of dynamically selected WSI-Fsets $\{\tilde{x}_0, \ldots, \tilde{x}_t, \ldots\}$. In the t-th step of the process, the MIL aggregator $M(\cdot)$ receives the current WSI-Fset $\tilde{x}_t$ and output the feature vector v_t, as well as the following slide-level embedding p_t by a projection head $f(\cdot)$. Then the slide-level embedding is sent to calculate the contrastive loss in Equ. (1). Meanwhile, an action s_{t+1} for selecting feature indexes of the next WSI-Fset $\tilde{x}_{t+1}$ will also be determined by the WSI-Fset proposal agent R by taking the feature vector v_t as inputs. Further, the WSI-Fset $\tilde{x}_{t+1}$ will be selected from the WSI-Fbag x with the generated feature indexes as described in Fig. 2 (c) and after that, it will be taken as the input in the next step. We take ABMIL [8] and CLAM [41] as the MIL aggregator $M(\cdot)$ to aggregate input features in WSI-Fset, while the projection head $f(\cdot)$ and the WSI-Fset proposal agent R are both recurrent networks as [33], such that they can exploit the information from all previous inputs by maintaining hidden state h_{t-1}^{R} and h_t^{f} within them respectively. It should be noted that our framework is not a RL framework, while we use the optimization strategy in RL to train our framwork. In the proposed MuRCL, the reinforcement learning is employed as an auxiliary task trained with the multi-instance learning for discriminative set construction. In the set construction procedure, the agent scans the WSI for several times to locate the discriminative features and calculates the reward in each step to update the agent for next decision making. Therefore, we borrow the idea of optimization approach in reinforcement learning for the set construction in our framework.

1) RL-Driven Selection: In each step of the RL-driven discriminative set construction, the WSI-Fset is selected from the WSI-Fbag with the feature indexes as the following procedure. To facilitate the agent to generate a spatial-consistent WSI-Fset proposal, we reorder the features in the input WSI-Fbag x according to their cluster labels, that is, features with the same cluster label are assigned adjacent indexes. After that, for features of each cluster C_k, we rearrange them along the coordinates of their corresponding patches. The reorder and rearranged WSI-Fbag is called Affinity map as shown in Fig. 2 (c). Then, we can compound the WSI-Fset from the Affinity map following the action s predicted by the agent R, in which we formulate the action s as a set of feature indexes for the rearranged clusters. Specifically, at the t-th step, with a feature index vector $s_t \in \mathbb{R}^{K*1}$ exported by the agent R at the $(t-1)$-th step, where element s_t^k denotes the feature indexes of the k-th rearranged cluster. Thus, for cluster C_k, we sample a feature sequence at the beginning of (s_t^k)-th feature, where the length of the sequence is equal to the sampling ratio multiplied by the number of cluster feature. Then, the WSI-Fset $\tilde{x}_t$ can be constructed by concatenating sequences of different clusters together.

2) Reward: The WSI-Fset proposal agent R is trained using the policy gradient method [42]. In the training procedure, how to design the reward function is important as it controls the direction in which the agent R is optimized [32]. In our framework, we propose to utilize the similarity between the positive pairs of WSI-Fsets as the reward to guide the agent to locate the informative features. Specifically, at the t-th step, the reward function to train the agent is designed as Equ. (2).

$$r_{i,j;t} = sim(p_{i;t-1}, p_{j;t-1}) - sim(p_{i;t}, p_{j;t}). \quad (2)$$

where p_i and p_j are derived from the corresponding input positive WSI-Fsets $\tilde{x}_i$ and $\tilde{x}_j$, respectively. By enlarging the cosine distance of these positive WSI-Fsets, the agent R is directed to pick out feature sets with larger divergence, thus promoting the MIL model to concentrate on latent aggregation knowledge by minimizing the CL loss.

3) Discriminative Set Mixup: To introduce more perturbation in MIL aggregator training, we also employ an effective feature augmentation strategy, named set-mixup, to increase the diversity of WSI-Fsets. For WSI-Fsets in a training batch, by mixing one WSI-Fset $\tilde{x}_l$ with the other $\tilde{x}_q$, the $\bar{x}_q$ is generated as a augmentation of $\tilde{x}_q$. It can be illustrated in Equ. (3), where λ is a coefficient sampled from a distribution

Authorized licensed use limited to: SICHUAN UNIVERSITY. Downloaded on February 23,2024 at 07:42:13 UTC from IEEE Xplore. Restrictions apply.

such as $\lambda \sim U(\alpha, 1.0)$, where α is set as 0.9 in our experiment.

$$\bar{x}_q = \lambda \tilde{x}_q + (1-\lambda)\tilde{x}_l. \tag{3}$$

The set-mixup operation is in favour of enhancing semantic concepts learning of the MIL aggregator by mixing WSI-Fsets which contain manifold WSI-level representations.

Algorithm 1 The Training Process of the MuRCL

Input: batch size N, constant τ, structure of $M(\cdot)$, $f(\cdot)$, R, hyperparameter α, γ.

for batch WSI-Fbags $\{x_n\}_{n=1}^{N}$ **do**
  for all $n \in \{1, \ldots, N\}$ **do**
    randomly select a WSI-Fsets $\tilde{x}_{n;0}$;
    $\lambda \sim U(\alpha, 1.0)$;
    $\tilde{x}_0 \sim \{\tilde{x}_{n;0}\}_{n=1}^{N} - \tilde{x}_{n;0}$;
    $\tilde{x}_{n;0} \leftarrow \lambda \tilde{x}_{n;0} + (1-\lambda)\tilde{x}_0$;
    $v_{n;0} = M(\tilde{x}_{n;0})$;
    $p_{n;0} = f(v_{n;0})$;
    $t = 1$
    repeat
      R selects a new WSI-Fset $\tilde{x}_{n;t}$ by taking $v_{n;t-1}$ as input state;
      $\lambda \sim U(\alpha, 1.0)$;
      $\tilde{x}_t \sim \{\tilde{x}_{n;t}\}_{n=1}^{N} - \tilde{x}_{n;t}$;
      $\tilde{x}_{n;t} \leftarrow \lambda \tilde{x}_{n;t} + (1-\lambda)\tilde{x}_t$;
      $v_{n;t} = M(\tilde{x}_{n;t})$;
      $p_{n;t} = f(v_{n;t})$;
      $t \leftarrow t + 1$
    until t=6
  end for
  for all $i \in \{1, \ldots, 2N\}$ and $j \in \{1, \ldots, 2N\}$, $t \in \{0, 1, 2, 3, 4, 5\}$ **do**
    $s_{i,j;t} = p_{i;t}^{T} p_{j;t} / (\| p_{i;t} \| \| p_{j;t} \|)$
  end for
  define $r(i, j; t) = s_{i,j;t-1} - s_{i,j;t}$
  define $l(i, j; t) = -log \frac{exp(s_{i,j;t}/\tau)}{\sum_{n=1}^{2N} \mathbb{1}(n \neq i) exp(s_{i,n;t}/\tau)}$
  update the $M(\cdot)$ and $f(\cdot)$ to minimize $L = \sum_{t=0}^{t=5} \frac{1}{2N} \sum_{n=1}^{N} [l(2n-1, 2n; t) + l(2n, 2n-1; t)]$
  update the R to maximize $R = \sum_{t=1}^{t=5} \gamma^{t-1} \frac{1}{N} \sum_{n=1}^{N} r(2n-1, 2n; t)$
end for
Return $M(\cdot)$, $f(\cdot)$, R

C. Training Scheme of RL-MIL

As shown in Fig. 2(a), our contrastive learning framework has two branches, where the parameters of the MIL aggregator $M(\cdot)$, the projection head $f(\cdot)$, and the agent R in two branches are shared. In a training batch, we first employ two randomly selected WSI-Fsets from the WSI-Fbag as the initialization of the positive pair. Then, the positive pair will be constructed by the agent trained with an RL strategy. For clear illustrations, we use the subscripts $(\cdot)_i$ and $(\cdot)_j$ to denote variables of two branches respectively and $\{(\cdot)_t\}_{t=0}^{T}$ to represent temporal sequence produced in each branch. T is set to five, indicating that the RNNs perform five iterations (steps) for a training batch. At the beginning, two randomly selected WSI-Fsets $\tilde{x}_{i;0}$ and $\tilde{x}_{j;0}$ from the same WSI-Fbag are regarded as the initialization positive pair and then two different feature vectors $v_{i;0}$ and $v_{j;0}$ are inferred by the $M(\cdot)$, respectively. Meanwhile, WSI-level feature embeddings $p_{i;0}$ and $p_{j;0}$ are calculated and forwarded to compute the CL loss $L_0 = L_{i,j;0}(p_{i;0}, p_{j;0})$ in which WSI-Fsets from different WSI-Fbags in the training batch are regarded as negative pairs. Next, the first iteration of R is performed and new positive pairs are proposed by taking $v_{i;0}$ and $v_{j;0}$ as initial states of R, respectively. With the new positive pair, the second inference iteration is conducted in the same way as the first one, followed by the second iteration of R. Over five iterations, the $f(\cdot)$ is processed synchronously with the R, and outputs of $f(\cdot)$ from two branches $\{p_{i;t}, p_{j;t}\}_{t=0}^{5}$ will be used to calculate the CL loss: $L_t = \sum_{t=0}^{5} L_{i,j}(p_{i;t}, p_{j;t})$ by Equ. (1). In addition, the agent R is trained simultaneously by maximizing discounted reward $r_{i,j} = \sum_{t=1}^{5} \gamma^{t-1} r_{i,j;t}(p_{i;t}, p_{j;t})$ where the γ is set to 0.1. Moreover, we derive $2N$ WSI-Fsets from N WSI-Fbags with two different samplings. For each positive pair, we treat the other $2(N-1)$ WSI-Fsets as negative examples follow [29]. In summary, the inference process can be clarified as Algorithm 1.

The training procedure includes three stages. In the first stage, the WSI-Fset proposal agent R is not included in the RL-MIL. Instead, we randomly sample a WSI-Fset from the input WSI-Fbag x in each step, and then train the MIL aggregator $M(\cdot)$ and projection head $f(\cdot)$ to minimize the contrastive loss in Equ. (1). This step is performed to ensure that the model can adapt to arbitrary input sequences. In the second stage, the MIL aggregator and the projection head are fixed. We randomly initialize the agent R and then train it following the training procedure described above. In the third stage, the MIL aggregator and the projection head are further fine-tuned with the fixed R obtained from the second step to enhance the performance of RL-MIL with the learned feature selection policy.

D. Framework Fine-Tuning and Inference

With the above multiple instance contrastive learning paradigm, the framework can deeply exploit the semantic relationship of different patches for slide-level WSI presentation. For final slide-level prediction, we further fine-tune the framework with labeled WSIs. In this phase, we change the final output dimension of $f(\cdot)$ from 128 to the class number to meet our downstream classification task. The fine-tuning training also consists of three stages as described in pre-training. Differently, we produce a confidence score by conducting a soft-max operation on the output of the $f(\cdot)$. Then the increment of the confidence score, *i.e.*, $\hat{r}_t = \hat{p}_t - \hat{p}_{t-1}$ ($\hat{(\cdot)}$ denotes the corresponding variables in fine-tuning) is taken as the reward of the WSI-Fset proposal agent R follows settings in [33], where $\hat{p}$ is the softmax prediction probability with the ground truth label.

The process of MuRCL in the testing phase is the same as the fine-tuning phase. At the beginning of the testing, with an input test WSI-Fbag, an initial state is provided by $M(\cdot)$ with random sampled WSI-Fset for agent R. Then, a strategically selected WSI-Fset will be determined by the agent R and processed by the MIL aggregator $M(\cdot)$ and a projector $f(\cdot)$ as in the fine-tuning phase. In this procedure, the agent with

Authorized licensed use limited to: SICHUAN UNIVERSITY. Downloaded on February 23,2024 at 07:42:13 UTC from IEEE Xplore. Restrictions apply.

the initial state iterative process the state vector, as the state of each iteration will be passed to the next iteration. The iteration is set to five times as the training and fine-tuning phase. We take the output of the agent R at the last iteration as the proposal of WSI-Fset and then take the output of the $f(\cdot)$ as the classification prediction. The contrastive loss can promote the model to enlarge the spatial distance among features of different categories and pull the distance between features of the same category closer. After fine-tuning, the model can make accurate class prediction by assigning the same category label to features with small spatial distance while different category label to features with large spatial distance. So as to distinguish the normal and cancer region. In addition, the RL agent can iteratively analyze features sampled from different positions to locate the discriminative features in a coarse to fine manner, which is similar to the repeat diagnosis process of doctors. Therefore, these components can be expected to improve the model.

IV. Experiments

A. Datasets

We report results on four benchmark clinical WSI datasets, including Camelyon16, TCGA-Lung, TCGA-Kidney [46] and TCGA-Esca that cover the cases of balanced/unbalanced and single/multiple class MIL problems. (1) **Camelyon16** is a widely used public dataset for metastasis detection in breast cancer, including 270 training WSIs and 129 test WSIs, and has been widely used for cancer/non-cancer classification and localization tasks [12]. After pre-processing, a total of about 2.7 million patches in average about 6881 patches per bag are obtained. (2) **TCGA-Lung** includes two subtype projects, *i.e.*, Lung Squamous Cell Carcinoma (TCGA-LUSC) and Lung Adenocarcinoma (TCGA-LUAD), for a total of 1041 diagnostic WSIs, including 529 LUAD slides and 512 LUSC slides, has been taken as benchmark in WSI subtype classification and survival analysis. After pre-processing, the mean number of patches extracted per slide is 11540. (3) **TCGA-Kidney** is consisted of three subtype sets, i.e., Kidney Chromophobe Renal Cell Carcinoma (TCGA-KICH), Kidney Renal Clear Cell Carcinoma (TCGA-KIRC) and Kidney Renal Papillary Cell Carcinoma (TCGA-KIRP), for a total of 734 diagnostic WSIs, including 92 KICH slides, 411 KIRC slides and 231 KIRP slides, which is suitable for multi-class subtype classification task. (4) **TCGA-Esca** is consisted of two subtype sets, for a total of 156 diagnostic WSIs with 90 squamous cell carcinoma and 66 adenocarcinoma, respectively.

We used 20× magnification WSIs in all experiments. The WSIs were cropped into patches with the size of 256 × 256 and we discarded the patches where the tissue region is lower than 35%. For Camelyon16, we randomly split the official training WSIs into 80% and 20% for training and validation, and test on the official test WSIs, following previous works [12]. For TCGA datasets, we randomly split the data in a ratio of 3:1:1 for training, validation, and testing, following [12].

B. Implementations Details and Evaluation Metrics

Similar to [41] and [17], each WSI patch is embedded into a 512-dimensional (d=512) feature vector with a pre-trained encoder. We employ the SimCLR [29] ResNet18 [39] encoder trained by [12] for the Camelyon16 and TCGA-Lung, while the TCGA-Kidney and TCGA-Esca are encoded by a ResNet18 pre-trained on ImageNet The WSI feature bag is first clustered into 10 categories using K-means [40] and the sampling ratio for each cluster C_k is set to $1024/u$ in the WSI-Fset construction, where the u denotes the number of features in the WSI bag. In addition, the batch size N is set to 128 in our experiments and the temperature τ is set as 1. In the first stage of the training procedure, we employed one Adam optimizer with two different initial learning rates of 1e-4 and 1e-5 to optimize the MIL aggregator $M(\cdot)$ and the projection head $f(\cdot)$, respectively. The weight decays were both set to 1e-5. For the second stage, we employed an Adam optimizer with an initial learning rate of 1e-5 to train the agent R. In the third stage, the MIL aggregator $M(\cdot)$ and the projection head $f(\cdot)$ are jointly optimized by an Adam optimizer, where the initial learning rates were set to 5e-5 and 1e-5, respectively, with the same weight decay of 1e-5. These three stages were trained with 100, 30 and 100 epochs, respectively. The optimizer setups for the fine-tuning phase are the same as the pre-training phase. All experiments were conducted with two RTX 3090 GPU cards. We take accuracy (ACC), area under the curve (AUC) scores, and F1 score to evaluate the classification performance, where the ACC is calculated with a threshold of 0.5 in all experiments. For AUC, we report the average one-versus-rest AUC (macro-averaged) for the multi-class classification case.

C. Comparisons With State-of-the-Arts

1) Experimental Settings: We present results of both binary-class and multi-class classification. The binary-class classification includes Camelyon16 cancer/non-cancer classification, TCGA-Lung cancer subtype classification and TCGA-Esca subtype classification. The multi-class classification includes TCGA-Kidney cancer subtype classification. We compare our framework with a set of strong baselines, including different advanced MIL models [4], [8], [41], [44]. As our MuRCL is model-agnostic to MIL aggregators, we evaluate our MuRCL with ABMIL [8] and CLAM_SB [41] as the MIL aggregator backbone (denoted as Ours(ABMIL) and Ours(CLAM_SB), respectively) to show the generalization and robustness of our framework. We first train our framework with the proposed MuRCL and then fine-tune it with labeled WSIs, while we also train other methods with the same labeled WSIs. We run each experiment five times with different seeds and reported the *average performance* with standard deviation.

2) Comparison Results: We first compare the performance of Camelyon16, which is the most popular WSI benchmark classification dataset. As shown in Table I and Table II, our frameworks with CLAM_SB and ABMIL aggregators outperform their corresponding baseline counterparts (*i.e.*, 5.5% for CLAM_SB and 3.9% for ABMIL on Camelyon16 in ACC), which shows that our MuRCL pre-train strategy is helpful to obtain aggregation knowledge. Overall, our framework achieves the best performance than other MIL methods, demonstrating the superiority of our MuRCL framework.

We can observe the similar trend for TCGA-Lung dataset. Also, our method outperforms all the other competing methods in all three metrics. It is observed that ABMIL shows poor

Authorized licensed use limited to: SICHUAN UNIVERSITY. Downloaded on February 23,2024 at 07:42:13 UTC from IEEE Xplore. Restrictions apply.

TABLE I
COMPARISONS OF OUR METHOD AND OTHER STATE-OF-THE-ART MIL METHODS ON CAMELYON16 AND TCGA-LUNG. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD

Dataset	Camelyon16			TCGA-Lung		
Method	ACC	AUC	F1	ACC	AUC	F1
MinMax [43]	87.29±2.33	90.39±0.56	81.42±3.34	84.40±1.44	91.79±1.04	83.36±1.46
Gated_ABMIL [8]	87.29±1.52	91.05±0.60	81.59±2.33	73.68±1.96	81.48±1.07	74.95±1.39
RNNMIL [4]	84.81±2.30	86.08±4.99	76.22±4.82	73.59±1.60	81.78±2.72	73.94±1.87
SetTransformer [44]	88.22±1.03	90.81±0.90	82.38±1.78	76.75±2.15	86.38±2.25	77.10±1.93
DeepAttnMIL [45]	88.68±1.16	92.73±0.74	83.19±1.95	75.31±0.55	81.41±1.00	74.72±1.20
DSMIL [12]	87.60±1.30	90.70±0.86	82.07±1.74	86.32±0.43	94.61±0.67	85.27±0.56
CLAM_MB [41]	86.28±2.31	90.58±1.97	80.88±3.30	87.63±1.35	95.18±0.84	86.97±1.58
CLAM_SB [41]	85.81±2.87	90.59±1.89	80.60±3.48	88.18±1.28	95.20±0.46	87.62±1.44
ABMIL [8]	86.32±2.18	89.10±2.55	79.69±3.46	78.37±1.91	86.62±1.04	76.66±1.88
Ours(CLAM_SB)	**91.32±1.92**	94.52±0.98	**88.02±2.67**	**89.19±0.72**	**96.37±0.32**	88.64±0.76
Ours(ABMIL)	90.23±2.28	**95.27±0.99**	86.42±2.65	88.90±1.30	95.81±0.46	**88.74±1.22**

TABLE II
COMPARISONS OF OUR METHOD AND OTHER STATE-OF-THE-ART MIL METHODS ON TCGA-KIDNEY AND TCGA-ESCA. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD

Dataset	TCGA-Kidney			TCGA-Esca		
Method	ACC	AUC	F1	ACC	AUC	F1
MinMax [43]	76.87±1.83	85.89±2.58	54.91±2.31	75.00±3.71	85.04±3.71	67.72±4.95
Gated_ABMIL [8]	82.31±1.83	93.08±0.64	77.13±2.19	72.05±3.98	82.34±2.95	58.97±5.67
RNNMIL [4]	74.83±4.15	88.31±3.54	69.18±4.51	73.85±6.26	81.83±5.83	68.24±8.99
SetTransformer [44]	83.81±1.17	94.16±0.73	80.25±1.78	75.77±3.76	86.75±1.70	73.65±5.28
DeepAttnMIL [45]	83.67±0.00	92.90±0.31	78.33±0.41	73.85±1.78	80.39±1.32	61.11±1.49
DSMIL [12]	84.90±1.12	94.68±0.14	81.27±1.51	83.97±4.70	93.15±2.22	78.01±6.96
CLAM_MB [41]	85.07±1.76	94.30±0.69	81.03±2.01	88.33±4.36	**94.14±2.30**	84.30±6.36
CLAM_SB [41]	84.25±1.42	94.74±0.60	79.65±2.04	87.18±3.81	93.55±2.20	82.76±5.79
ABMIL [8]	78.91±2.71	92.77±1.12	60.68±9.08	74.49±2.69	84.63±2.42	68.64±3.10
Ours(CLAM_SB)	84.49±1.57	**95.73±0.38**	80.61±2.30	**88.72±1.40**	93.75±1.29	**86.08±1.64**
Ours(ABMIL)	**85.85±0.79**	95.68±0.45	**82.38±1.35**	83.08±2.92	89.08±0.70	79.96±4.87

TABLE III
P-VALUES OF OUR METHOD AND OTHER STATE-OF-THE-ART MIL METHODS ON THREE BENCHMARK DATASETS

Dataset	Camelyon16		TCGA-Lung		TCGA-Kidney	
Method	Ours(CLAM_SB)	Ours(ABMIL)	Ours(CLAM_SB)	Ours(ABMIL)	Ours(CLAM_SB)	Ours(ABMIL)
DSMIL [12]	4.7e-4	1.5e-4	8.3e-4	1.4e-2	1.4e-2	2.6e-3
CLAM_MB [41]	1.1e-4	5.17e-5	2.0e-4	3.4e-2	3.7e-3	4.3e-4
CLAM_SB [41]	1.1e-4	5.0e-5	6.6e-5	2.4e-2	2.3e-2	4.4e-3

performance on TCGA-Lung due to the limited representation capability of the extracted patch features. However, with our proposed reinforcement contrastive learning framework, the performance of ABMIL is largely improved. We believe that it is because our MuRCL can guide the backbone MIL network to concentrate on more informative features based on the generic knowledge learned from raw data. Also, our method is compatible with different MIL backbones, suggesting similar conclusions as the Camelyon16 dataset.

Different from the previous datasets, TCGA-Kidney consists of three categories WSIs. Moreover, the TCGA-Kidney is unbalanced distributed in cancer subtypes, which makes it more difficult for accurate classification. In this case, our MuRCL is also applicable to multi-class problems with an unbalanced dataset and consistently improves the corresponding baseline counterparts in all metrics. The experimental results on this dataset demonstrate the generalization and robustness of our proposed framework for various sites.

To further demonstrate the robustness of the proposed method, we also conduct analysis experiments on TCGA-Esca which consists of fewer data (156 slides) than the other three datasets, namely Camelyon16 (399 slides), TCGA-Lung (1041 slides) and TCGA-Kidney (734 slides). As shown in Table II, our method achieves comparable performance to other state-of-the-art methods. It shows that under limited training data scenario, the reinforcement contrastive learning framework can still learn representative features for the classification task.

As the tumor/normal classification is more challenging than the subtype classification. Our method shows more advantages over other MILs on the Camelyon16 than the TCGA datasets, showing that our framework can extract and learn discriminative features to handle more challenging WSI analysis tasks. Consider that there are few public datasets for tumor/normal classification, to demonstrate the robustness of our method, we conduct additional experiments on three TCGA datasets to valid the effectiveness of our method. Moreover, we conduct statistics test of different methods to demonstrate that the proposed method significantly outperforms other approaches. The AUC P-values for DSMIL, CLAM_MB and CLAM_SB compared to Ours(CLAM_SB) and Ours(ABMIL) are presented

Authorized licensed use limited to: SICHUAN UNIVERSITY. Downloaded on February 23,2024 at 07:42:13 UTC from IEEE Xplore. Restrictions apply.

TABLE IV
RESULTS OF DIFFERENT ABLATION MODELS AFTER FINE-TUNING

Dataset					Camelyon16			TCGA-Lung			TCGA-Kidney		
Method	**CL**	**RL**	**MIL**	RL	ACC	AUC	F1	ACC	AUC	F1	ACC	AUC	F1
MIL aggregator (ABMIL)													
M_A					86.98±1.73	91.02±0.77	81.76±2.14	73.11±1.45	81.98±0.59	74.98±0.50	82.45±2.56	92.75±3.40	72.65±9.51
M_B				✓	**90.39±1.05**	94.88±0.66	85.89±1.69	86.22±1.74	94.88±1.57	85.98±1.41	84.76±1.33	95.20±0.45	80.67±1.35
M_{max}	✓		✓		83.37±11.47	88.53±9.22	77.93±12.23	69.83±2.55	77.41±2.19	71.58±2.80	71.56±2.92	86.44±2.26	54.17±6.99
M_{min}	✓		✓		83.37±11.88	88.34±9.43	77.30±16.02	70.12±1.90	77.85±1.88	72.75±1.44	74.69±2.20	87.16±2.27	65.48±3.79
M_C	✓		✓		86.51±1.26	91.01±0.74	80.20±1.60	82.97±1.61	89.88±0.91	81.66±1.53	82.86±1.17	93.88±0.48	78.03±1.26
M_D	✓	✓	✓		87.44±0.76	91.05±0.59	82.02±1.33	83.35±4.54	91.27±2.97	82.56±4.62	83.40±1.53	94.41±0.51	79.20±1.86
M_E	✓	✓	✓	✓	90.23±2.28	**95.27±0.99**	**86.42±2.65**	**88.90±1.30**	**95.81±0.46**	**88.74±1.22**	**85.85±0.79**	**95.68±0.45**	**82.38±1.35**
MIL aggregator (CLAM_SB)													
M_A					86.82±1.10	91.04±0.53	81.88±1.15	88.42±0.70	95.54±0.51	88.03±0.72	83.81±0.90	94.90±0.56	79.66±1.09
M_B				✓	90.85±2.21	93.91±0.50	87.11±3.25	89.19±1.20	96.26±0.49	88.58±1.22	84.49±1.55	**95.95±0.68**	79.70±1.43
M_{max}	✓		✓		86.01±1.87	90.90±1.28	80.97±2.32	87.20±2.36	94.58±0.90	86.37±2.96	85.27±1.58	95.43±0.44	81.46±1.90
M_{min}	✓		✓		86.12±1.95	90.86±1.25	81.08±2.30	87.42±1.76	94.56±0.82	86.70±2.06	85.27±1.44	95.35±0.47	81.41±1.91
M_C	✓		✓		86.05±1.30	91.09±0.38	80.96±1.39	87.66±1.07	94.44±0.80	86.96±1.34	85.03±1.77	95.24±0.55	81.25±2.43
M_D	✓	✓	✓		87.44±0.90	91.36±0.83	82.19±1.32	87.66±1.15	94.46±0.77	87.14±1.38	**86.12±1.02**	95.25±0.17	**83.18±1.40**
M_E	✓	✓	✓	✓	**91.32±1.92**	**94.52±0.98**	**88.02±2.67**	**89.19±0.72**	**96.37±0.32**	**88.64±0.76**	84.49±1.57	95.73±0.38	80.61±2.30

in Table VI. Most P-values are smaller than 0.05 and the proposed method significantly outperforms other approaches.

D. Analysis of Our Framework

1) Ablation Study: We conduct an ablation study to show the effectiveness of our proposed multi-instance contrastive learning mechanism and reinforcement learning module for selecting discriminative set. We first compare the features learned by MuRCL to several self-supervised learning baselines. Table IV presents the ablation experiment results of different models after the supervised fine-tuning. The reported results are also the average performance of five runs with different seeds.

Model A (M_A) is the vanilla MIL aggregator with randomly sampled WSI-Fset and Model B (M_B) is the supervised MIL aggregator using the RL-agent to generate the WSI-Fset. Model min (M_{min}) is the vanilla MIL aggregator with WSI-Fset sampled with minimum cosine similarity of the extracted features. Specifically, we random sample ten feature sets from the bag, and choose a pair of set with the smallest cosine similarity as the positive pair. Similarly, Model max (M_{max}) is the vanilla MIL aggregator with WSI-Fset sampled with maximum cosine similarity of the extracted features. Specifically, we random sample ten feature sets from the bag, and choose a pair of set with the largest cosine similarity as the positive pair. It is observed that M_B achieves a much large improvement than M_A, showing the effectiveness of RL-based discriminative feature selection strategy. In Model C (M_C), we use a vanilla contrastive learning strategy to train the model, where the positive/negative pairs are constructed with randomly selected features. M_C achieves better results than M_A, showing the effectiveness of the contrastive learning pre-training. For Model D (M_D), we use the proposed MuRCL framework to train the framework but do not utilize RL-agent to reconstruct WSI-Fset during fine-tuning (*i.e.*, using randomly selected WSI-Fset during fine-tuning). M_D achieves better performance than M_C on TCGA-Kidney dataset and Camelyon16 and comparable performance with M_C on TCGA-Lung datasets, showing that RL-based WSI-Fset construction can improve the quality of positive pairs for more efficient CL training. In Table IV, our final model (M_E) achieves the best performance than all ablation models in most cases, demonstrating the capability of our proposed whole framework. We also conduct experiments of the proposed method without set mixup (Ours w/o set-mixup) and show the results in the Table VII. It demonstrates that the set mixup can effectively improve the model performance.

TABLE V
RESULTS OF DIFFERENT NUMBER OF ITERATIONS FOR RUNNING RL DURING TRAINING

Iter(T)	ACC	AUC	F1
Camelyon16			
T=2	86.20±1.27	90.96±0.96	80.86±1.69
T=4	86.67±2.35	93.15±0.71	81.78±3.19
T=5	**91.32±1.92**	**94.52±0.98**	**88.02±2.67**
Lung			
T=2	87.18±0.71	95.41±0.31	86.67±0.92
T=4	88.04±0.76	95.43±0.62	87.49±0.68
T=5	**89.19±0.72**	**96.37±0.32**	**88.64±0.76**
Kidney			
T=2	83.27±0.78	95.01±0.40	79.42±1.45
T=4	83.81±0.57	95.24±0.49	79.83±0.98
T=5	**84.49±1.57**	**95.73±0.38**	**80.61±2.30**

TABLE VI
RESULTS OF MULTI-TASK EXPERIMENTS WITH 10% LABELS ON CAMELYON16

Method	ACC	AUC	F1
CLAM	85.58±0.62	90.01±1.05	81.51±0.77
Multi-task	87.75±1.01	90.66±1.50	82.08±1.49
MuRCL	**88.06±3.01**	**92.48±0.90**	**84.66±3.42**

TABLE VII
ABLATION EXPERIMENTS FOR SET-MIXUP ON CAMELYON16

Method	ACC	AUC	F1
Ours w/o set-mixup	88.06±2.23	94.07±0.72	83.48±2.72
Ours	**91.32±1.92**	**94.52±0.98**	**88.02±2.67**

2) Analysis of RL-MIL Design: From the above comparisons in Table IV, we can see that the RL-based WSI-Fset construction is not only suitable for self-supervised MIL training (M_D vs. M_C), but also has a strong promoting effect on fully-supervised MIL training (M_B vs. M_A). We believe it is because the reinforcement module could effectively collect information from previous actions and make better decisions, which also demonstrates that constructing a discriminative

Authorized licensed use limited to: SICHUAN UNIVERSITY. Downloaded on February 23,2024 at 07:42:13 UTC from IEEE Xplore. Restrictions apply.

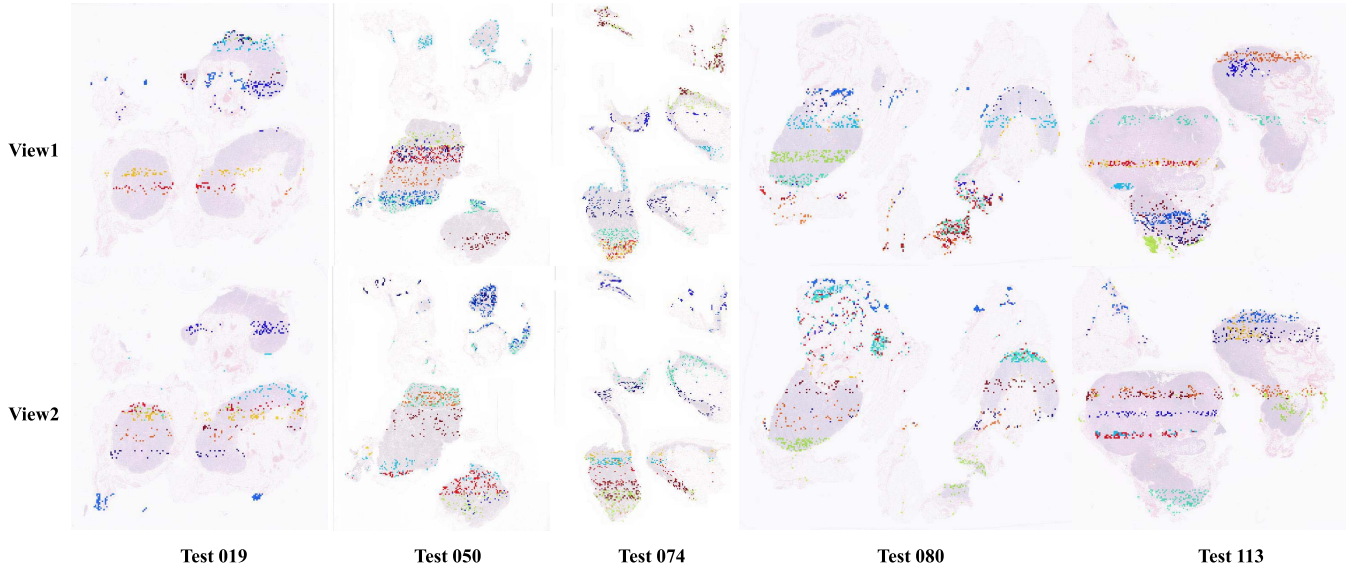

Fig. 3. Visualizations of features in a pair of discriminative sets (View1 and View2) from Camelyon16. The features selected in each clustering category are highlighted in an individual color.

TABLE VIII
LINEAR EVALUATION EXPERIMENTS OF OURS

Method	CL	RL	MIL	RL	Camelyon16 ACC	Camelyon16 AUC	Camelyon16 F1	TCGA-Lung ACC	TCGA-Lung AUC	TCGA-Lung F1	TCGA-Kidney ACC	TCGA-Kidney AUC	TCGA-Kidney F1
MIL aggregator (ABMIL)													
M_C	✓				62.02±0.00	88.62±0.94	0.00±0.00	54.07±0.80	59.80±0.13	51.71±0.76	58.37±0.27	77.47±0.35	29.57±0.12
M_D	✓	✓	✓		62.02±0.00	**89.60±0.88**	0.00±0.00	50.72±0.00	71.71±0.74	0.00±0.00	55.78±0.00	75.04±1.00	23.87±0.00
M_E	✓	✓	✓	✓	**82.02±0.76**	79.49±0.96	**73.64±1.01**	**80.77±2.06**	**88.72±1.01**	**79.61±1.95**	**82.86±1.09**	**93.92±0.75**	**78.92±1.40**
MIL aggregator (CLAM_SB)													
M_C	✓				75.19±0.49	80.95±0.53	51.50±1.42	72.15±0.36	82.84±0.30	70.40±0.47	75.65±0.67	90.34±0.18	71.02±0.95
M_D	✓	✓	✓		79.07±0.49	79.17±0.14	65.10±1.17	73.30±0.36	83.83±0.24	71.73±0.46	78.78±0.51	92.84±0.11	74.50±0.79
M_E	✓	✓	✓	✓	**80.89±1.05**	**84.68±0.61**	**69.94±1.19**	**83.35±1.53**	**92.40±0.51**	**82.49±1.98**	**82.31±0.75**	**94.97±0.35**	**77.90±0.65**

TABLE IX
THE INFLUENCE OF DIFFERENT FEATURE REPRESENTATIONS ON CAMELYON16 AND TCGA-LUNG

Method	CL	RL	MIL	RL	Camelyon16 ACC	Camelyon16 AUC	Camelyon16 F1	TCGA-Lung ACC	TCGA-Lung AUC	TCGA-Lung F1
Encoder					ImageNet ResNet18			ImageNet ResNet18		
CLAM_SB					75.16±4.89	70.32±4.14	61.90±5.08	84.71±1.18	93.10±0.65	84.23±1.27
M_B				✓	77.21±4.30	73.58±2.04	67.40±4.45	86.70±1.49	93.79±0.70	86.35±1.63
M_E	✓	✓	✓	✓	**79.07±2.95**	**78.56±1.33**	**69.90±2.68**	**87.08±1.31**	**94.43±0.34**	**86.72±1.02**
Encoder					SimCLR ResNet18			SimCLR ResNet18		
CLAM_SB					86.82±1.10	91.04±0.53	81.88±1.15	88.42±0.70	95.54±0.51	88.03±0.72
M_B				✓	90.85±2.21	93.91±0.50	87.11±3.25	89.19±1.20	96.26±0.49	88.58±1.22
M_E	✓	✓	✓	✓	**91.32±1.92**	**94.52±0.98**	**88.02±2.67**	**89.19±0.72**	**96.37±0.32**	**88.64±0.76**

feature set from WSI is very beneficial to multi-instance learning. Also, we study the number of iterations for running RL during training as shown in Table V. As shown in the table, more iteration leads to better performance. We believe this is because the agent can select the most discriminative features from more iterations by more sufficient comparison. However, each iteration needs to store intermediate variable parameters, so it consumes a lot of memory. Limited by the memory consumption, we set the number of iterations as 5 in our experiments. In the future, the influence of more iterations can be explored with more GPU memory.

3) Analysis of CL Design: As presented in Table IV, the M_D pretrained with reinforcement contrastive learning achieves better performance than M_A, which demonstrates that the contrastive learning pretrain plays an important role in the MIL. Moreover, we apply different encoding networks include ResNet18 pretrained on ImageNet (ImageNet ResNet18) and ResNet18 pretrained by SimCLR (SimCLR ResNet18) from [12] to extract features from the Camelyon16 and TCGA-Lung datasets. Since the [12] did not provide a pre-trained SimCLR ResNet18 on the TCGA-Kidney, we employ the ResNet50 pretrained on ImageNet (ImageNet ResNet50) and ImageNet ResNet18 to extract features from the TCGA-Kidney. As shown in the Table IX and Table X, with different feature representations, the proposed reinforcement contrastive learning can still have a positive effect on improving the classification performance (compare the CLAM_SB to M_B and M_E). In addition, comparing results

Authorized licensed use limited to: SICHUAN UNIVERSITY. Downloaded on February 23,2024 at 07:42:13 UTC from IEEE Xplore. Restrictions apply.

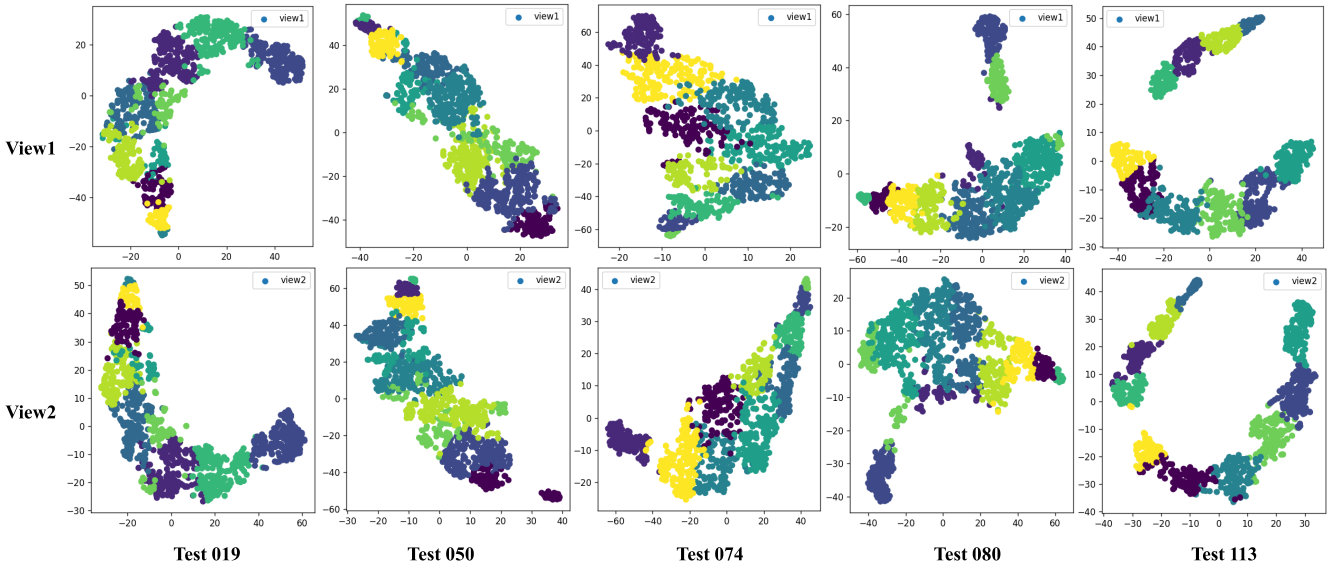

Fig. 4. Visualizations of t-sne in a pair of discriminative sets from Camelyon16. The features selected in each clustering category are highlighted in an individual color.

TABLE X
THE INFLUENCE OF DIFFERENT FEATURE REPRESENTATIONS ON TCGA-KIDENY

Dataset					TCGA-Kidney		
Method	**CL**	**RL**	**MIL**	RL	ACC	AUC	F1
Encoder					ImageNet ResNet50		
CLAM_SB					85.26±0.84	95.34±0.44	80.44±2.53
M_B				✓	84.76±1.33	**96.07±0.31**	80.44±2.53
M_E	✓	✓	✓	✓	**86.26±0.57**	95.81±0.52	**83.05±0.48**
Encoder					ImageNet ResNet18		
CLAM_SB					83.81±0.90	94.90±0.56	79.66±1.09
M_B				✓	84.49±1.55	**95.95±0.68**	79.70±1.43
M_E	✓	✓	✓	✓	**84.49±1.57**	95.73±0.38	**80.61±2.30**

of ImageNet ResNet18 and SimCLR ResNet18, we can find that the encoder (SimCLR ResNet18) pretrained with training data can help to promote the performance of all methods.

4) WSI Discriminative Set Visualization: We visualize the position of the selected WSI-Fset of the pretrained MuRCL model to analyze the effectiveness of the proposed MuRCL. Fig. 3 visualizes the location of two selected discriminative sets and Fig. 4 shows the t-sne [47] visualization of its features on WSI examples from Camelyon16. Due to the lack of pixel-level annotation, the RL agent needs to scan the WSI regions to confirm if there is a cancer patch. However, due to uncertainty errors of the model, it is not reasonable to determine the label of the WSI only by one patch predicted as metastatic tumor. The model needs to collect sufficient features from all patches to get a more precise prediction. Therefore, the model need to look at the WSI region in a spectific order determined by the RL agent to discover the discriminative features according to the change of feature distribution. Moreover, we present the prediction heatmap generated from the model trained in Pretrained and Fine-tune manners in Fig. 5, where the probability scores are in a range from 0 to 1 (blue to red). In addition, we also compare the heatmap with the tumor region ground truth, which is presented in the yellow curve in Fig. 5. Comparing these maps with the cancer area annotated by doctors, it is easy to find that the model and the doctor tend to have the same region of interest. This phenome could explain why the model can make accurate diagnosis. On the one hand, we measure the overlap between the different discriminative sets constructed by the model on all the slides of the Camelyon16 and get the result of 10.24±9.17% and 19.75±23.66% for tumor slides and normal slides, respectively. In Fig. 3, different colors represent selected features of different clusters. As illustrated in Section III-B1, we put the features of each cluster in order according to the coordinates of these features in the WSI, and then select features to construct the discriminative set from each cluster. Since the features of the same cluster have continuous coordinates in the WSI, the selected features in the same cluster will appear as a scan-line pattern in the visualizations. As shown in Fig. 3, the selected two views have discrimination between each other but contain the same semantic information with the WSI, which is helpful for the model to learn latent class representation in the contrastive learning. Also, our MuRCL can help to localize the critical tumor patches without pixel-level annotations (see zoomed-in patches), contributing to better WSI classification performance.

V. DISCUSSION

We present a novel multi-instance reinforcement contrastive learning for WSI classification. Our method has demonstrated significant promotion over previous MIL methods on representative benchmark WSI datasets. The pivotal technical innovation is a novel set-based self-supervised MIL training paradigm derived from patch-based contrastive learning. We are motivated to alleviate the overfitting problem in WSI MIL as mentioned before, by mining more relationships among different patches using reinforcement contrastive learning.

Besides, the proposed MuRCL framework can also be regarded as a strong regularization during supervised training. Since our MuRCL is a two-stage procedure, to evaluate its regularization effect, we conduct a one-stage multi-task learning experiment on Camelyon16 dataset with 10% labeled training data. In this experiment, we also add a supervised task in the MuRCL framework (see Fig. 2 (a)). Specifically, we add a classifier following p_i and calculate a classification

Authorized licensed use limited to: SICHUAN UNIVERSITY. Downloaded on February 23,2024 at 07:42:13 UTC from IEEE Xplore. Restrictions apply.

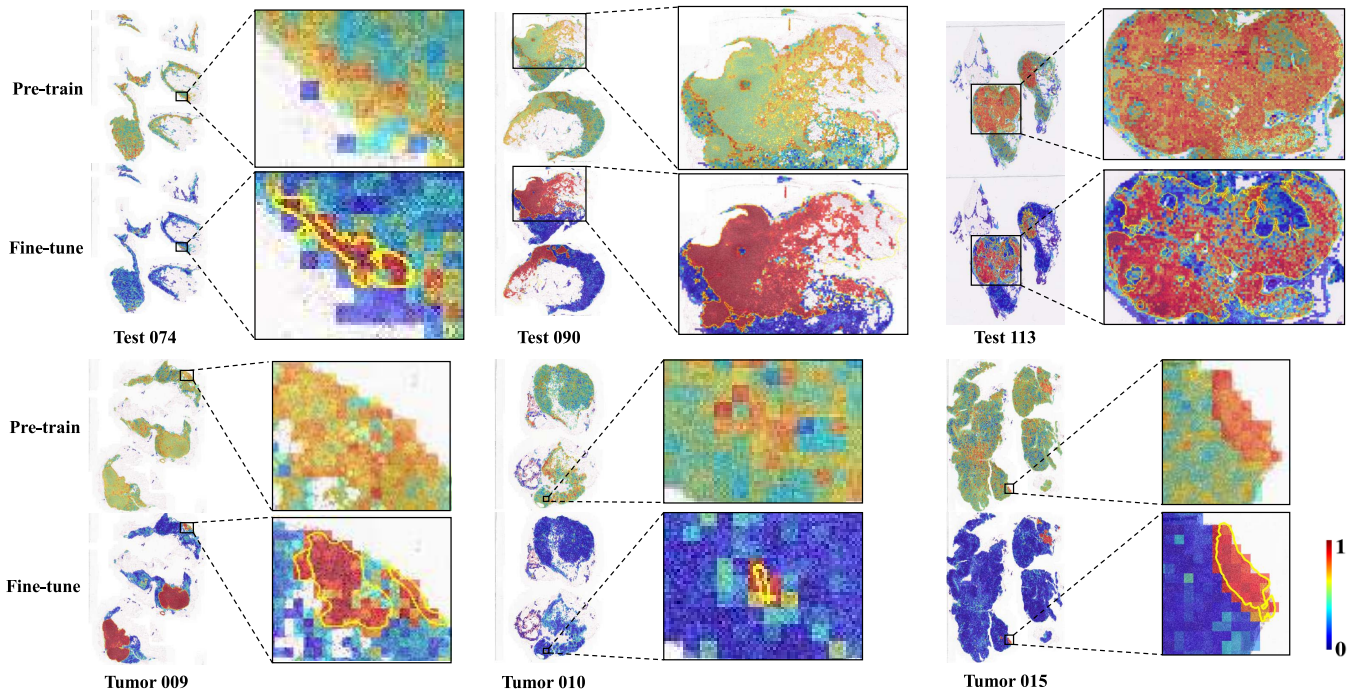

Fig. 5. Visualizations of self-supervised (Pre-train) heatmap and supervised (Fine-tune) heatmap from Camelyon16. These heatmaps denote the attention score of each patch in the WSI inferred by a self-supervised trained model and a supervised trained model, where the yellow curve represents the localization annotation.

loss. Results are shown in Table X. Although the performance of the multi-task model is inferior to our fine-tuned MuRCL, it achieves better performance than the MIL baseline, showing the regularization effect of our proposed reinforcement contrastive learning framework.

To show the effectiveness of our proposed multi-instance contrastive learning paradigm, we also evaluate the learned representations of our MuRCL using the linear evaluation protocol [30], [48], [49], where a linear classifier is trained on top of a frozen encoder network, and the test results are used as a proxy for representation quality. For linear evaluation, we use an Adam optimizer with a learning rate of 1e-4 and the training process is performed in 40 epochs. The results are depicted in Table VIII. It is observed that the M_C also gains superior performance than M_A and M_B under the linear evaluation. Moreover, M_B gets better results than M_A in most cases, further demonstrating the effectiveness of RL-based WSI-Fset construction in CL training.

We propose to make use of reinforcement learning to build discriminative feature sets for positive/negative pairs in self-supervised MIL contrastive learning. In reinforcement learning, how to design the action and reward is important and we leverage the proposed affinity map and contrastive learning loss to design these key components. We will investigate different strategies to further improve the effectiveness of the proposed RL-agent, such as the optimization of the reward function and the proposal mechanism of the agent. For example, since a WSI contains millions of patches, it is a time-consuming and memory-consuming process to learn the features together with the aggregator/classifier for WSIs. Therefore, it may be impractical to fine-tune the feature extractor together with the aggregator/classifier using GPU with limited memory. We can use the set-based multi-instance learning, which constructs discriminative patch sets, to train the model to alleviate the GPU consumption. However, this approach may significantly increase the training time consumption. We will focus on learning the feature extraction together with the aggregator/classificatory more efficiently in our future work.

In terms of practicality, we believe our work has the potential for clinical applications where annotations are limited and provides a new insight for set-based self-supervised research. Moreover, the WSI analysis is a tedious, laborious and time-consuming process due to the large scale of the image. Our method can automatically help doctors to make faster and better diagnosis, which can alleviate the workload of doctors. In addition, our concept of self-supervised MIL training can be generalized to data with higher dimensions (*e.g.* 3D data or videos) and more MIL setting, which could be our future works.

VI. Conclusion

In this paper, we propose a novel and efficient MIL framework for WSI classification, which also holds great potential for other MIL problems. To combat the overfitting in MIL tasks caused by insufficient training data, we formulate the learning process as a self-training procedure to learn latent relationships among different instances and thus regularize the whole framework. Furthermore, the self-training procedure is designed as a set-based CL process and the reinforcement learning is incorporated to help select discriminative positive pair for CL. Finally, the pretrained model is finetuned with slide-level labels for final prediction. Both quantitative and qualitative evaluations demonstrate the superiority of the proposed method on three independent datasets.

Authorized licensed use limited to: SICHUAN UNIVERSITY. Downloaded on February 23,2024 at 07:42:13 UTC from IEEE Xplore. Restrictions apply.

REFERENCES

[1] F. Xing, Y. Xie, H. Su, F. Liu, and L. Yang, "Deep learning in microscopy image analysis: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 10, pp. 4550–4568, Oct. 2018.

[2] G. Litjens et al., "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017.

[3] L. Hou, D. Samaras, T. M. Kurc, Y. Gao, J. E. Davis, and J. H. Saltz, "Patch-based convolutional neural network for whole slide tissue image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2424–2433.

[4] G. Campanella et al., "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images," *Nature Med.*, vol. 25, no. 8, pp. 1301–1309, Aug. 2020.

[5] P.-H.-C. Chen et al., "An augmented reality microscope with real-time artificial intelligence integration for cancer diagnosis," *Nature Med.*, vol. 25, no. 9, pp. 1453–1457, Sep. 2019.

[6] H. Zhang et al., "DTFD-MIL: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification," 2022, *arXiv:2203.12081*.

[7] J. Feng and Z.-H. Zhou, "Deep MIML network," in *Proc. AAAI Conf. Artif. Intell.*, 2017, vol. 31, no. 1, pp. 1884–1890.

[8] M. Ilse, J. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 2127–2136.

[9] X. Wang, Y. Yan, P. Tang, X. Bai, and W. Liu, "Revisiting multiple instance neural networks," *Pattern Recognit.*, vol. 74, pp. 15–24, Feb. 2018.

[10] O. Dehaene, A. Camara, O. Moindrot, A. de Lavergne, and P. Courtiol, "Self-supervision closes the gap between weak and strong supervision in histology," 2020, *arXiv:2012.03583*.

[11] M. Y. Lu, R. J. Chen, J. Wang, D. Dillon, and F. Mahmood, "Semi-supervised histology classification using deep multiple instance learning and contrastive predictive coding," 2019, *arXiv:1910.10825*.

[12] B. Li, Y. Li, and K. W. Eliceiri, "Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14318–14328.

[13] G. Quellec, G. Cazuguel, B. Cochener, and M. Lamard, "Multiple-instance learning for medical image and video analysis," *IEEE Rev. Biomed. Eng.*, vol. 10, pp. 213–234, 2017.

[14] P. Chikontwe, M. Kim, S. J. Nam, H. Go, and S. H. Park, "Multiple instance learning with center embeddings for histopathology classification," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2020, pp. 519–528.

[15] R. J. Chen et al., "Multimodal co-attention transformer for survival prediction in gigapixel whole slide images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 4015–4025.

[16] Y. Sharma, A. Shrivastava, L. Ehsan, C. A. Moskaluk, S. Syed, and D. E. Brown, "Cluster-to-conquer: A framework for end-to-end multi-instance learning for whole slide image classification," 2021, *arXiv:2103.10626*.

[17] Z. Shao et al., "TransMIL: Transformer based correlated multiple instance learning for whole slide image classification," 2021, *arXiv:2106.00908*.

[18] C. L. Srinidhi, S. W. Kim, F.-D. Chen, and A. L. Martel, "Self-supervised driven consistency training for annotation efficient histopathology image analysis," 2021, *arXiv:2102.03897*.

[19] L. Chen, P. Bentley, K. Mori, K. Misawa, M. Fujiwara, and D. Rueckert, "Self-supervised learning for medical image analysis using image context restoration," *Med. Image Anal.*, vol. 58, Dec. 2019, Art. no. 101539.

[20] X. Li, M. Jia, M. T. Islam, L. Yu, and L. Xing, "Self-supervised feature learning via exploiting multi-modal data for retinal disease diagnosis," *IEEE Trans. Med. Imag.*, vol. 39, no. 12, pp. 4023–4033, Dec. 2020.

[21] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 11, pp. 4037–4058, Nov. 2021.

[22] 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.

[23] A. Taleb, C. Lippert, T. Klein, and M. Nabi, "Multimodal self-supervised learning for medical image analysis," in *Proc. Int. Conf. Inf. Process. Med. Imag.* Springer, 2021, pp. 661–673.

[24] W. Bai et al., "Self-supervised learning for cardiac MR image segmentation by anatomical position prediction," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2019, pp. 541–549.

[25] M. Blendowski, H. Nickisch, and M. P. Heinrich, "How to learn from unlabeled volume data: Self-supervised 3d context feature learning," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2019, pp. 649–657.

[26] S. Azizi et al., "Big self-supervised models advance medical image classification," 2021, *arXiv:2101.05224*.

[27] B. Cao, H. Zhang, N. Wang, X. Gao, and D. Shen, "Auto-GAN: Self-supervised collaborative learning for medical image synthesis," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 10486–10493.

[28] F. Mahmood, R. Chen, and N. J. Durr, "Unsupervised reverse domain adaptation for synthetic medical images via adversarial training," *IEEE Trans. Med. Imaging*, vol. 37, no. 12, pp. 2572–2581, Dec. 2018.

[29] 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.

[30] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9729–9738.

[31] K. Chaitanya, E. Erdil, N. Karani, and E. Konukoglu, "Contrastive learning of global and local features for medical image segmentation with limited annotations," 2020, *arXiv:2006.10511*.

[32] J. Hong et al., "Reinforced attention for few-shot learning and beyond," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 913–923.

[33] Y. Wang, K. Lv, R. Huang, S. Song, L. Yang, and G. Huang, "Glance and focus: A dynamic approach to reducing spatial redundancy in image classification," 2020, *arXiv:2010.05300*.

[34] Y. Wang, Z. Chen, H. Jiang, S. Song, Y. Han, and G. Huang, "Adaptive focus for efficient video recognition," 2021, *arXiv:2105.03245*.

[35] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "AutoAugment: Learning augmentation strategies from data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 113–123.

[36] K. Wang, B. Kang, J. Shao, and J. Feng, "Improving generalization in reinforcement learning with mixture regularization," 2020, *arXiv:2010.10814*.

[37] R. Raileanu, M. Goldstein, D. Yarats, I. Kostrikov, and R. Fergus, "Automatic data augmentation for generalization in reinforcement learning," in *Proc. Workshop Inductive Biases, Invariances Gen., Int. Conf. Mach. Learn.*, 2020, pp. 5402–5415.

[38] M. Laskin, A. Srinivas, and P. Abbeel, "Curl: Contrastive unsupervised representations for reinforcement learning," in *Proc. ICML*, 2020, pp. 5639–5650.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[40] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. IT-28, no. 2, pp. 129–137, Mar. 1982.

[41] M. Y. Lu, D. F. K. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, and F. Mahmood, "Data-efficient and weakly supervised computational pathology on whole-slide images," *Nature Biomed. Eng.*, vol. 5, no. 6, pp. 555–570, Mar. 2021.

[42] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017, *arXiv:1707.06347*.

[43] P. Courtiol, E. W. Tramel, M. Sanselme, and G. Wainrib, "Classification and disease localization in histopathology using only global labels: A weakly-supervised approach," 2018, *arXiv:1802.02212*.

[44] J. Lee, Y. Lee, J. Kim, A. Kosiorek, S. Choi, and Y. W. Teh, "Set transformer: A framework for attention-based permutation-invariant neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 3744–3753.

[45] J. Yao, X. Zhu, J. Jonnagaddala, N. Hawkins, and J. Huang, "Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks," *Med. Image Anal.*, vol. 65, Oct. 2020, Art. no. 101789.

[46] K. Tomczak, P. Czerwińska, and M. Wiznerowicz, "The cancer genome atlas (TCGA): An immeasurable source of knowledge," *Contemp. Oncol.*, vol. 19, no. 1A, p. A68, 2015.

[47] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.

[48] P. Bachman, R. D. Hjelm, and W. Buchwalter, "Learning representations by maximizing mutual information across views," 2019, *arXiv:1906.00910*.

[49] O. Henaff, "Data-efficient image recognition with contrastive predictive coding," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 4182–4192.

Authorized licensed use limited to: SICHUAN UNIVERSITY. Downloaded on February 23,2024 at 07:42:13 UTC from IEEE Xplore. Restrictions apply.