Dual Space Multiple Instance Representative Learning for Medical Image Classification

Xiaoxian Zhang[1]
zhangxiaoxian@cqu.edu.cn
Sheng Huang[1,*]
huangsheng@cqu.edu.cn
Yi Zhang[1]
zhangyii@cqu.edu.cn
Mingchen Gao[2]
mgao8@buffalo.edu
Chen Liu[3]
liuchen@aifmri.com
Xiaohong Zhang[1]
xhongz@cqu.edu.cn

[1] School of Big Data & Software Engineering
Chongqing University
Chongqing, P.R.China

[2] Department of Computer Science and Engineering
University at Buffalo, SUNY
New York, USA

[3] Department of Radiology
First Affiliated Hospital of Army Medical University
Chongqing, P.R.China

Abstract

Medical image classification plays a vital role in AI-aided medical diagnosis and is often addressed as a Multiple Instance Learning (MIL) issue (i.e., each sample is a bag of instances). For medical images, the disease area or the discriminative area is usually smaller than the whole tissue. In other words, most of the instances in a bag are irrelevant and could interfere with the bag label inference. To address this issue, we add an instance representative selection process before MIL and propose a novel MIL approach named Dual Space Multiple Instance Representative Learning (DSMIRL). DSMIRL consists of two core steps, Adaptive Instance Representative Selection (AIRS) and Multiple Instance Representative Learning (MIRL). In AIRS, the instances in the same bag are grouped into different sub-bags via clustering, and only one sub-bag is selected as the final collection of instance representatives by ranking the maximum instance predictions of sub-bags, thus adaptively filtering out the irrelevant instances. In MIRL, we perform aggregations on the selected instance representatives in label and feature spaces to further exploit the complementary information of the two spaces. Finally, these two steps are iteratively conducted in each iteration to optimize all modules of DSMIRL progressively. Extensive experiments on five standard MIL benchmarks and two medical image datasets demonstrate the promising performance of DSMIRL over the state-of-the-art MIL approaches.

1 Introduction

Medical image classification as a fundamental task of medical image analysis plays an important role in AI-aided medical diagnosis. Medical image classification has achieved remarkable progress with the help of deep neural networks [15, 30, 43]. However, such a significant

advancement highly relies on large-scale medical data labeled in a fine-grained manner (i.e., pixel-wise labels or patch-wise labels). Labeling is labor-intensive, time-consuming process and often needs fruitful expert knowledge, which is unrealistic in many real-world scenarios. Medical image classification based on Multiple Instance Learning (MIL) approaches is gaining popularity [3, 28, 31, 48], requiring only coarse-grained labels for model optimization. In MIL [27], each sample is a bag consisting of multiple instances.

In general, existing MIL algorithms can be divided into two categories, namely, instance-level methods [4, 16, 41] and embedding-level methods [7, 12, 13, 24, 40], based on different aggregating spaces. Instance-level methods conduct aggregations in the label space with the max-pooling or mean-pooling operation. Embedding-level methods conduct aggregations in the feature space and leverage the aggregated feature to infer the bag label. Among them, attention-based methods [7, 13, 40] are typical embedding-level methods.

In the medical image, the diseased instances tend to be a small fraction of all instances (i.e., typically less than 20% [20]), leading to a strong imbalance between positive (disease) and negative (normal) instances. Superabundant negative samples may weaken the identification of the positive instances with the conventional aggregating functions, such as max-pooling and mean-pooling. Although attention-based MIL approaches can alleviate this issue via modeling the relations between instances [19, 20, 33]. However, these methods still consider irrelevant instances in MIL model optimization and thus cannot thoroughly eliminate the effect of these irrelevant instances [21].

In this paper, we provide an intuitive method for addressing the issue mentioned above. Our idea is to add an instance representative selection process before MIL. This process is used to filter out irrelevant instances while selecting relevant instances as a sub-bag to predict bag-level labels. Aggregation is only performed in the selected instances to eliminate the effects of the irrelevant instances during MIL as many as possible. To implement this idea, we elaborate a novel MIL approach named Dual Space Multiple Instance Representative Learning (DSMIRL), which consists of two core steps, namely Adaptive Instance Representative Selection (AIRS) and Multiple Instance Representative Learning (MIRL). In AIRS, the instances in the same bag are divided into different sub-bags by measuring the similarities of their features extracted by the pre-trained model. One sub-bag is selected as the final collection of instance representatives by ranking the maximum instance predictions of sub-bags. In MIRL, we perform aggregations on the selected instance representatives in label and feature spaces simultaneously for accomplishing MIL. Five standard MIL benchmarks and two medical image datasets are employed to evaluate our method. Extensive results validate the effectiveness of our method. Our contributions are summarized as follows,

- We introduce a novel idea for MIL, which adds an additional instance selection step to eliminate the effects of the irrelevant instances in model optimization before performing MIL. Based on this idea, we propose a novel MIL approach named DSMIRL. Extensive experimental results on several MIL and medical image datasets demonstrate its superiority over baselines.

- We elaborate a simple but effective clustering-based instance representative selection method, which adaptively selects the most relevant instance as the instance representation based on the features and predictions in each iteration.

- We conduct an attention-based aggregation in the feature space and mean-pooling in the label space as a dual space aggregation strategy to fully exploit the complementary information of the feature and label spaces.

2 Related Work

Multiple instance learning (MIL) [9] is a weakly supervised learning framework that requires only coarse-grained label instead of elaborated fine-grained annotation. MIL is widely used in multiple machine learning tasks, such as semantic segmentation [36, 44], object detection [17, 34, 37], scene classification [18, 38], and text categorization [47, 49]. Likewise, MIL-based medical image classification is gradually becoming a trend [5, 26, 32, 35, 41, 42].

In general, MIL algorithms can be divided into two groups, namely, instance-level algorithms [1, 4, 6, 16, 22, 46] and embedding-level algorithms [7, 13, 19, 20, 33, 40, 45]. For the former, bag-level prediction is usually obtained through max-pooling or mean-pooling in the label space. By contrast, the latter aggregates instance features into a bag-level representation and then learns a bag-level classifier for bag-level prediction based on the bag-level representation. Although instance-level algorithms tend to learn less discriminative sample information, they also take a higher overfitting risk and are prone to interference from negative instances. Instance-level algorithms are empirically proven to be inferior to embedding-level counterparts in performance [33].

Most of the embedding-level algorithms are attention-based, and they differ in the manner they generate attention scores. For instance, Ilse et al. propose an attention-based aggregation operator, which gives each instance additional contribution information through trainable attention weight network [19]. Li et al. introduce non-local attention to model the instance-to-instance and instance-to-bag relations. The weight of each instance is obtained by calculating its similarity to the key instance [20]. Shao et al. adopt a self-attention mechanism in Transformer to focus on the pairwise correlation between each instance within a bag [33]. These methods of aggregating in feature space tend to retain more intrinsic information about the samples for label inference. However, since there are only a small number of diseased regions in the medical image, a large amount of information is redundant and even interferes with sample identification.

In contrast to the approaches mentioned above, we propose to add an instance representative selection process to filter out irrelevant instances before MIL aggregation. Therefore, we devise a simple but effective AIRS module to adaptively select instance representatives and design a MIRL module to exploit complementary information in feature and label spaces.

3 Methodology

3.1 Preliminary and Overview

Multiple Instance Learning (MIL) is a typical weakly supervised learning technique that has been widely used in medical image classification. In this paper, we formulate the medical image classification task as a MIL problem. The medical images or image patches from the same patient are considered a bag B. Thereby the dataset can be denoted as $\mathcal{D} = \{(B_1, \mathcal{Y}_1), (B_2, \mathcal{Y}_2), \cdots, (B_N, \mathcal{Y}_N)\}$, where $\mathcal{Y}$ is the bag-level label and N is the number of bags. Each bag consists of a sequence of instances (e.g., images or image patches). A bag can be represented as $B = \{x_1, x_2, \cdots, x_n\}$, where $\mathcal{Y} = \{y_1, y_2, \cdots, y_n\}$ is the corresponding instance-level labels and n is the number of instances in a bag. In general, the number of instances is variable in different bags. In MIL, the bag-level label $\mathcal{Y}$ is available while the instance-level label y is unknown. Our task is to infer the bag label via aggregating instance information. In this paper, we propose Dual Space Multiple Instance Representative Learning (DSMIRL) to solve this task, which consists of three modules, namely Feature Learning

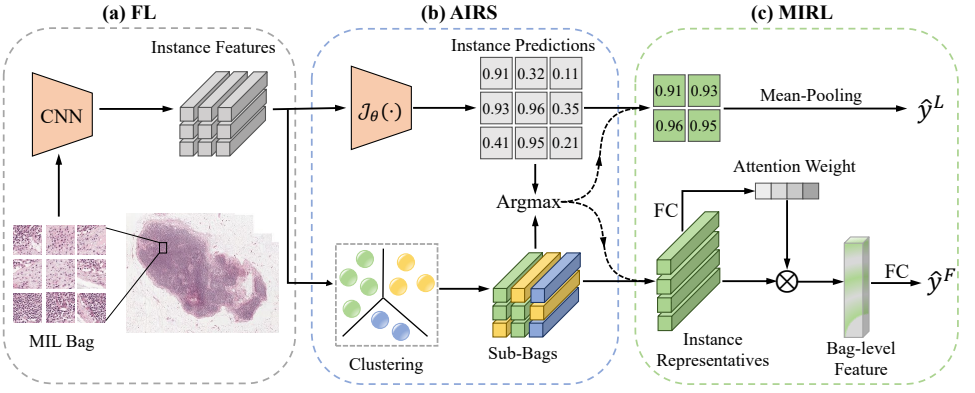

Figure 1: The framework of Dual Space Multiple Instance Representative Learning (DSMIRL), which consists of Feature Learning (FL), Adaptive Instance Representative Selection (AIRS), and Multiple Instance Representative Learning (MIRL). In this approach, AIRS is employed to filter out the irrelevant instances based on the instance features extracted by FL, and then MIRL conducts the instance aggregation in both feature and label spaces to exploit their complementary information for better accomplishing the MIL task.

(FL), Adaptive Instance Representative Selection (AIRS), and Multiple Instance Representative Learning (MIRL). DSMIRL accomplishes the MIL task in two steps. The first step is to adaptively select the instance representatives with AIRS in an unsupervised manner based on the instance features extracted by FL.

3.2 Feature Learning

Similar to the previous work [33], we adopt ResNet50 [12] as the basic feature learning network on medical image datasets. We stack two fully connected layers to reduce the dimension of the instance feature to 512. The feature extraction procedure is denoted as $f_i = \mathcal{F}_\varphi(x_i)$, where f_i is a 512-dimensional feature vector and φ is the parameters of the feature learning network. Therefore, the instances of each bag can be represented as $\mathcal{F}_\varphi(B) = \{f_1, f_2, \cdots, f_n\}$.

3.3 Adaptive Instance Representative Selection

The disease area is often smaller than the whole tissue in medical images. This finding implies that negative instances hold a much larger portion than positive instances. These superabundant negative instances can easily interfere with the follow-up instance aggregation in MIL. We elaborate an adaptive unsupervised instance selection step before performing the final MIL to avoid this issue. We first divide each bag into multiple clusters by clustering methods. Then, each cluster is considered a sub-bag and scored by the maximum value of instance scores obtained by instance label prediction network $\mathcal{J}_\theta(\cdot)$. Finally, the instance selection task degenerates into a sub-bag score ranking problem. The instances of the sub-bag with the highest score are selected as the instance representatives of the corresponding bag. Notably, such a process is conducted in each iteration. Our method is an end-to-end learning framework, and AIRS will progressively guide the optimization of feature learning and the MIRL module by adaptively reserving the relevant instances in each iteration.

Different instances of the same disease area are similar in cells and tissues, so we can utilize the similarity between instances to model instance relationships in a bag. Since the instances are unlabeled, we group the bag into clusters by clustering according to the instance relationships $M = \pi\left(\{f_i\}_{i=1}^{n}\right)$, where $\pi(\cdot)$ can be any clustering method. This will be empirically discussed in the experiment part (see Section 4.3.2). M is a indicator matrix encoding the clustering results and the i-th column of M indicates the incidence relation between instances and the i-th cluster, e.g.,

$$M_{:i}^{K} = \begin{matrix} f_1 & f_2 & f_3 & f_4 & \cdots & f_n \\ [\,1 & 0 & 1 & 0 & \cdots & 0\,] \end{matrix},$$

where K is the number of clusters. In our method, each cluster is deemed as a sub-bag b_k, and the sub-bag can be retrieved based on the indicator matrix,

$$b_k \leftarrow \Phi(B, M_{:k}),\ \text{s.t.} \bigcup_{k=1}^{K} b_k = B \text{ and } \bigcap_{k=1}^{K} b_k = \oslash, \tag{1}$$

where b_k is the k-th sub-bag and $\Phi(\cdot,\cdot)$ is an sub-bag retrieval operation based on the bag and the indicator matrix.

The instances in the same sub-bag are highly similar. Therefore, each sub-bag represents different types of instances, and the instance selection process is translated as the sub-bag selection process. We expect to filter out the instances that possess a low response to the final bag label inference. To achieve this goal, we score each sub-bag by using the maximum of instance-label predictions in each sub-bag, which can be seen as the latent responses of instances (sub-bag) to the final task,

$$p_k = \max(\{\hat{y}_i\}_{f_i \in b_k}) = \max(\{\mathcal{J}_\theta(f_i)\}_{f_i \in b_k}), \tag{2}$$

where p_k is the score of the k-th sub-bag b_k. $\mathcal{J}_\theta(\cdot)$ is the instance label prediction network, and θ is its associated learnable parameters. The sub-bag who owns the maximum score is picked up as the collection of instance representatives with respect to a bag $b_{\hat{k}} \leftarrow \arg\max_{b_k \in B} p_k$. Finally, only selected instance representatives will take part in the instance aggregation, effectively eliminating the interferences of the irrelevant instances in a bag.

3.4 Multiple Instance Representative Learning

In contrast to conventional MIL approaches, which apply aggregation on all instances in a bag, DSMIRL only performs aggregation on instances selected by AIRS, which is also referred to as instance representatives. In addition, most MIL approaches aggregate instances in a single space, mainly in feature space. However, the information of instances encoded in different spaces may be complementary because they reflect the same bag from different perspectives. In this section, we introduce a novel dual space instance aggregation strategy to fully exploit the information of feature and label spaces as well as incorporate the merits of the two different instance aggregations. In this strategy, we introduce an attention module to aggregate instance representatives in feature space for yielding bag-level feature while accomplishing instance aggregation in the label space through mean pooling.

Aggregation in Feature Space: After obtaining the instance representative embeddings, we employ attention-based MIL pooling to aggregate instance features. Similar to the work [19], the attention module consists of two fully connected layers, which can learn different weights for each instance adaptively. The features of instance representatives are weighted and summed to produce a final bag-level feature: $\hat{f} = \sum_{f_i \in b_{\hat{k}}} a_i f_i$, where a_i is the

learnable weight corresponding to the i-th instance representative,

$$a_i = \frac{\exp\left\{u^T \tanh(v f_i^T)\right\}}{\sum_{f_j \in b_{\hat{k}}} \exp\left\{u^T \tanh(v f_j^T)\right\}}, \text{ s.t. } f_i, f_j \in b_{\hat{k}}, \tag{3}$$

where u and v are the parameters of two fully connected layers. We introduce a one-layer neural network as classifier to infer the bag label with this bag-level feature $\hat{\mathcal{Y}}^F = \mathcal{Q}_\psi(\hat{f})$, where $\mathcal{Q}_\psi(\cdot)$ is the classifier while ψ is its parameters.

Aggregation in Label Space: With regard to the instance aggregation in label space, we directly conduct the mean-pooling on the label predictions of instance representatives for achieving another bag-level label prediction,

$$\hat{\mathcal{Y}}^L = \frac{1}{|b_{\hat{k}}|} \sum_{x_i \in b_{\hat{k}}} \hat{y}_i. \tag{4}$$

The label predictions of instance representatives can reflect whether the AIRS module works well. Instances in the same sub-bag should have similar label predictions (e.g., all are normal or disease). Hence, we add their average prediction results as additional supplementary information to guide the AIRS module.

3.5 Model optimization and inference

The Cross-Entropy function $\mathcal{H}(\cdot,\cdot)$ is leveraged to measure the discrepancy between label predictions and ground-truths, and the overall loss $\mathcal{L}$ is denoted as follows:

$$\mathcal{L} = \frac{1}{2} \sum_{\mathcal{D}} \{\mathcal{H}(\mathcal{Y}, \hat{\mathcal{Y}}^F) + \mathcal{H}(\mathcal{Y}, \hat{\mathcal{Y}}^L)\}. \tag{5}$$

The DSMIRL model can be solved as the following optimization problem with Back Propagation $\{\hat{\varphi}, \hat{\theta}, \hat{u}, \hat{v}, \hat{\psi}\} \leftarrow \arg\min_{\varphi,\theta,u,v,\psi} \mathcal{L}$. We add the instance label space information in the training stage to help the AIRS module complete instance representative selection. In the inference stage, we only adopt the instance feature space information to infer the bag label.

4 Experiments and Results

4.1 Experimental Setup

We comprehensively evaluate DSMIRL on five MIL benchmarks (Elephant, Fox, Tiger, Musk1, and Musk2) and two medical image datasets (Camelyon16 and Pneumonia CT). The details about the datasets and implementation are described in the supplementary material.

4.2 Performance comparison with exiting methods

Table 1 shows the accuracies of our method and the compared approaches on MIL benchmarks. We can observe that our method achieves state-of-the-art performance on all datasets. For instance, the accuracy gains of our model over the suboptimal traditional method miGraph on Musk1, Musk2, Fox, Tiger, and Elephant are 7.7%, 5.7%, 16.9%, 12.0%, and 6.6%, respectively. In addition, our method achieves 1.9%, 2.6%, 4.0%, 5.0%, and 0.6% performance gains on the five datasets, respectively, compared with the second-best deep learning-based algorithm DSMIL. These phenomena validate the effectiveness of our method.

The phenomena similar to the ones on MIL benchmarks can be observed in Table 2. Compared with recent state-of-the-art approaches, our method exhibits the best performance on two medical image datasets and achieves considerable advantages. On Camelyon16, our model achieves 2.0%, 0.8%, and 2.7% performance gains over runner-up in accuracy, AUC, and F1-score, respectively. Moreover, DSMIRL achieves improved results than directly performing attention operations. For example, DSMIRL outperforms AttMIL by 2.7%, 1.6%, and 2.7% in terms of accuracy, AUC, and F1-score, respectively. DSMIRL still performs best on Pneumonia CT. DSMIL [20] is the second best-performed approach. The performance gains of DSMIRL over it are 1.9%, 1.1%, and 2.3%in accuracy, AUC, and F1-score, respectively. These results clearly validate that our proposed AIRS and MIRL modules effectively suppress the interference of negative instances and make full use of dual space information to further improve the performance of model.

Methods	Datasets				
	Musk1	Musk2	Fox	Tiger	Elephant
mi-SVM [2]	.874±N/A	.836±N/A	.582±N/A	.784±N/A	.822±N/A
mi-Graph [49]	.889±.03	.903±.04	.620±.04	.860±.04	.869±.04
MI-Kernel [11]	.880±.03	.893±.02	.603±.03	.842±.01	.843±.02
mi-Net [39]	.889±.04	.858±.05	.613±.04	.824±.03	.858±.04
AttMIL [19]	.892±.04	.858±.05	.615±.04	.839±.02	.868±.02
DSMIL [20]	.932±.02	.930±.02	.729±.02	.869±.01	.925±.01
DSMIRL(ours)	**.966±.05**	**.960±.04**	**.785±.07**	**.921±.07**	**.935±.05**

Table 1: The accuracy of different MIL approaches on MIL benchmarks (mean±std) with the previous method results taken from [19, 20]. Experiments use the same training setting as [19]. The highest accuracy is in bold, and the second-best accuracy is underlined.

Methods	Camelyon16			Pneumonia CT		
	Accuracy	AUC	F1-score	Accuracy	AUC	F1-score
Max-pooling	.864±.02	.920±.03	.821±.03	.835±.05	.895±.05	.834±.05
Mean-pooling	.859±.03	.917±.03	.836±.03	.849±.01	.903±.01	.851±.01
AttMIL [19]	.862±.02	.937±.01	.839±.01	.897±.02	.957±.01	.895±.01
DSMIL [20]	.862±.02	.930±.01	.839±.02	.911±.01	.956±.01	.907±.01
CLAM-SB [24]	.869±.03	.936±.02	.819±.04	.903±.01	.958±.01	.900±.01
CLAM-MB [24]	.852±.04	.934±.01	.807±.06	.885±.02	.947±.02	.886±.02
TransMIL [33]	.857±.03	.945±.02	.800±.06	.866±.05	.943±.02	.876±.05
DSMIRL(ours)	**.889±.01**	**.953±.01**	**.866±.02**	**.930±.01**	**.967±.01**	**.930±.01**

Table 2: The performance of different MIL approaches on Camelyon16 and Pneumonia CT. The highest performance is in bold, and the second-best performance is underlined.

4.3 Ablation Study

4.3.1 Effects of different modules

In this part, we conduct the ablation study on medical image datasets to quantify the effects of different modules in our method on MIL performance. Table 3 reports the performances

of our model when we add or remove the modules. The baseline of DSMIRL is the simple MIL approach, which aggregates instances in the label space with mean-pooling. On the Camelyon16, if we introduce the instance selection module (i.e., AIRS) in the baseline, the baseline is boosted by 2.1%, 2.9%, and 1.3% in accuracy, AUC, and F1-score, respectively. Moreover, if adding the dual-space instance aggregation strategy (i.e., MIRL), this model can be further improved by 0.9%, 0.7%, and 1.7% in the same evaluation metrics. On the Pneumonia CT, we observed a similar phenomenon. When adding the AIRS module, the model can obtain 5.1%, 4.9%, and 5.1% performance gains in accuracy, AUC, and F1-score, respectively. When further adding the MIRL module, the model can achieve performance improvements on all evaluation metrics. These results clearly verify that conducting the proper instance selection before instance aggregation can significantly benefit MIL, and performing the instance aggregations in feature and label spaces is able to further boost MIL.

Datasets	Module			
	Baseline	✓	✓	✓
	AIRS	×	✓	✓
	MIRL	×	×	✓
Camelyon16	Accuracy	.859±.03	.880±.02	**.889±.01**
	AUC	.917±.03	.946±.01	**.953±.01**
	F1-score	.836±.03	.849±.02	**.866±.02**
Pneumonia CT	Accuracy	.849±.01	.900±.03	**.930±.01**
	AUC	.903±.01	.952±.02	**.967±.01**
	F1-score	.851±.01	.902±.03	**.930±.01**

Table 3: Module analysis of DSMIRL on medical image datasets (including Camelyon16 and Pneumonia CT). DSMIRL=baseline+AIRS+MIRL.

Methods	Accuracy	AUC	F1-score
K-means [23]	**.894±.02**	.947±.01	.865±.02
Mean-shift [8]	.879±.02	.948±.01	.862±.02
DBSCAN [10]	.882±.02	.949±.01	.857±.02
Hierarchical Clustering [25]	.889±.01	.950±.01	**.869±.02**
Spectral Clustering [29]	.889±.01	**.953±.01**	.866±.02

Table 4: Results on Camelyon16. The best ones are in bold, and the second-best ones are underlined.

4.3.2 Discussion of different clustering strategies

In the AIRS module, we propose to stratify instances using a clustering method. We discuss the performances of DSMIRL when we adopt different clustering approaches in AIRS. Here, we adopt five clustering approaches, namely K-means [23], Mean-shift [8], DBSCAN [10], Hierarchical Clustering [25], and Spectral Clustering [29]. Tables 4 report the results of DSMIRL on Camelyon16 using different clustering methods in the AIRS module. Comprehensively speaking, Spectral Clustering performs the best among all five clustering methods. Hence, we choose Spectral Clustering as the clustering method of our AIRS module.

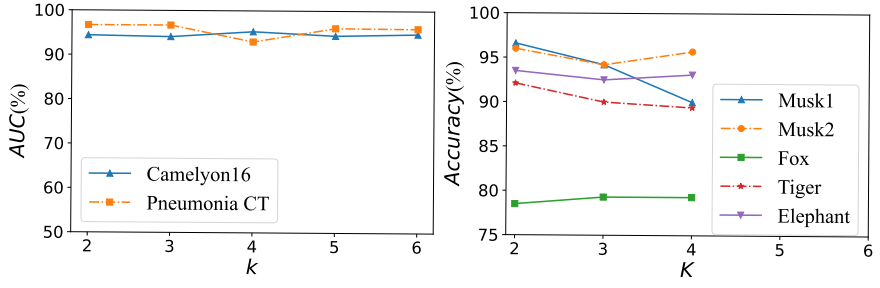

Figure 2: Effect of K on the performances of DSMIRL. We discuss it on two medical image datasets (Camelyon16 and Pneumonia CT). We only discuss three cases where K is 2, 3, and 4 on MIL benchmarks. Please refer to the supplementary material for details.

4.3.3 Effects of the number of clusters

The number of clusters K is an important parameter for controlling the scale of instance representatives. A larger K implies a small size of sub-bag and also a small scale of preserved instance representatives. If $K = 1$, it means that all instances are deemed as instance representatives. If $K = n$, then MIRL degenerates as a special max-pooling-based MIL approach, where n is the number of instances in a bag. We discuss the relationship between the value of K and DSMIRL performance. As shown in Figure 2, the performance of the model does not increase as K increases. The Camelyon16 shows a trend of first rising and then falling. When $K = 4$, AUC reaches its peak. Pneumonia CT reports the opposite direction, with the AUC maintaining optimal performance when $K = 2$. Considering the performance of the DSMIRL under the five metrics, we empirically set K to 4 and 2 on Camelyon16 and Pneumonia CT, respectively. We only discuss three cases of $K = 2$, 3, and 4 on MIL benchmarks. Combined with the performance of five small-scale datasets, we empirically set $K = 2$ on MIL benchmarks. For details, see the supplementary material.

4.4 Representative Instance Visualization

Considering that the Pneumonia CT does not provide detailed annotation information, we only conduct visualization experiments on the Camelyon16. Figure 3(c) shows the results of our selection, where the patches covered in light blue are the selected instance representatives, and the dark blue line in Figure 3(b) marks the disease area annotated by the domain experts. From Figure 3(c), we can see that our proposed model always enables highlighting the diseased patches as instance representatives, and thereby boosts the performance.

5 Conclusions

We present a novel MIL approach named DSMIRL for medical image classification. DSMIRL introduces an instance representative selection process before MIL for filtering out irrelevant instances, which may interfere with instance aggregation. It employs clustering on instance features to select a cluster with a maximum prediction as instance representatives for a bag. Then, the instance representatives are aggregated in the feature space and label space to accomplish the final MIL task. The dual-space instance aggregation strategy can further improve model performance by exploiting the complementary information of feature space and label space. Experimental results on five MIL benchmarks and two medical image datasets demonstrate the superiority of our method over the recent state-of-the-art approaches, and

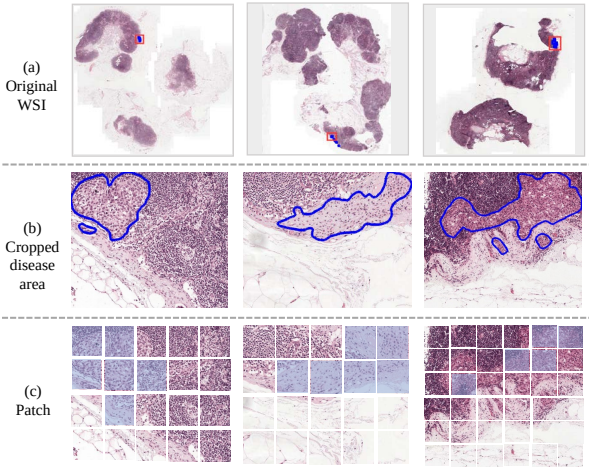

Figure 3: The visualization of the selected instance representatives. (a) The original WSI images from the Camelyon16 dataset. (b) The image is cropped from the entire WSI at 20x magnification, where the dark blue line is the disease area marked by experts. The disease area is from the area in the red rectangle above. (c) Patches covered in light blue are the selected instance representatives.

the ablation study validates our claimed contributions one by one. In the future, we plan to integrate clustering and representative selection operations into a network, making the cohesion of modules more natural and the network architecture more compact.

Acknowledgements This work was supported in part by the National Natural Science Foundation of China under Grant 62176030, in part by the Natural Science Foundation of Chongqing under Grant cstc2021jcyj-msxmX0568.

References

[1] Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. Support vector machines for multiple-instance learning. In *NeurIPS*, volume 2, page 7, 2002.

[2] Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. Support vector machines for multiple-instance learning. *NeurIPS*, pages 577–584, 2003.

[3] Qi Bi, Shuang Yu, Wei Ji, Cheng Bian, Lijun Gong, Hanruo Liu, Kai Ma, and Yefeng Zheng. Local-global dual perception based deep multiple instance learning for retinal disease classification. In *MICCAI*, pages 55–64. Springer, 2021.

[4] Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 25(8):1301–1309, 2019.

[5] Junhua Chen, Haiyan Zeng, Chong Zhang, Zhenwei Shi, Andre Dekker, Leonard Wee, and Inigo Bermejo. Lung cancer diagnosis using deep attention-based multiple instance learning and radiomics. *Medical Physics*, 49(5):3134–3143, 2022.

[6] Philip Chikontwe, Meejeong Kim, Soo Jeong Nam, Heounjeong Go, and Sang Hyun Park. Multiple instance learning with center embeddings for histopathology classification. In *MICCAI*, pages 519–528. Springer, 2020.

[7] Philip Chikontwe, Miguel Luna, Myeongkyun Kang, Kyung Soo Hong, June Hong Ahn, and Sang Hyun Park. Dual attention multiple instance learning with unsupervised complementary loss for covid-19 screening. *Medical Image Analysis*, 72:102105, 2021.

[8] Dorin Comaniciu and Peter Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, 2002.

[9] Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1-2): 31–71, 1997.

[10] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, volume 96, pages 226–231, 1996.

[11] Thomas Gärtner, Peter A Flach, Adam Kowalczyk, and Alexander J Smola. Multi-instance kernels. In *ICML*, volume 2, page 7, 2002.

[12] Zhongyi Han, Benzheng Wei, Yanfei Hong, Tianyang Li, Jinyu Cong, Xue Zhu, Haifeng Wei, and Wei Zhang. Accurate screening of covid-19 using attention-based deep 3d multiple instance learning. *IEEE Transactions on Medical Imaging*, 39(8): 2584–2594, 2020.

[13] Noriaki Hashimoto, Daisuke Fukushima, Ryoichi Koga, Yusuke Takagi, Kaho Ko, Kei Kohno, Masato Nakaguro, Shigeo Nakamura, Hidekata Hontani, and Ichiro Takeuchi. Multi-scale domain-adversarial multiple-instance cnn for cancer subtype classification with unannotated histopathological images. In *CVPR*, pages 3852–3861, 2020.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[15] Xingxin He, Ying Deng, Leyuan Fang, and Qinghua Peng. Multi-modal retinal image classification with modality-specific attention network. *IEEE Transactions on Medical Imaging*, 40(6):1591–1602, 2021.

[16] Le Hou, Dimitris Samaras, Tahsin M Kurc, Yi Gao, James E Davis, and Joel H Saltz. Patch-based convolutional neural network for whole slide tissue image classification. In *CVPR*, pages 2424–2433, 2016.

[17] Fang Huang, Jinqing Qi, Huchuan Lu, Lihe Zhang, and Xiang Ruan. Salient object detection via multiple instance learning. *IEEE Transactions on Image Processing*, 26 (4):1911–1922, 2017.

[18] Guixin Huang, Sheng Huang, Luwen Huangfu, and Dan Yang. Weakly supervised patch label inference network with image pyramid for pavement diseases recognition in the wild. In *ICASSP*, pages 7978–7982. IEEE, 2021.

[19] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *ICML*, pages 2127–2136. PMLR, 2018.

[20] Bin Li, Yin Li, and Kevin W Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *CVPR*, pages 14318–14328, 2021.

[21] Hang Li, Fan Yang, Xiaohan Xing, Yu Zhao, Jun Zhang, Yueping Liu, Mengxue Han, Junzhou Huang, Liansheng Wang, and Jianhua Yao. Multi-modal multi-instance learning using weakly correlated histopathological images and tabular clinical information. In *MICCAI*, pages 529–539. Springer, 2021.

[22] Guoqing Liu, Jianxin Wu, and Zhi-Hua Zhou. Key instance detection in multi-instance learning. In *ACML*, pages 253–268. PMLR, 2012.

[23] Stuart Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

[24] Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5(6):555–570, 2021.

[25] Alena Lukasová. Hierarchical agglomerative clustering procedure. *Pattern Recognition*, 11(5-6):365–381, 1979.

[26] Siyamalan Manivannan, Caroline Cobb, Stephen Burgess, and Emanuele Trucco. Subcategory classifiers for multiple-instance learning and its application to retinal nerve fiber layer visibility classification. *IEEE Transactions on Medical Imaging*, 36(5): 1140–1150, 2017.

[27] Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. *NeurIPS*, 10, 1997.

[28] Andriy Myronenko, Ziyue Xu, Dong Yang, Holger R Roth, and Daguang Xu. Accounting for dependencies in deep learning based multiple instance learning for whole slide imaging. In *MICCAI*, pages 329–338. Springer, 2021.

[29] Andrew Y Ng, Michael I Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *NeurIPS*, pages 849–856, 2002.

[30] Yuanyuan Peng, Weifang Zhu, Zhongyue Chen, Meng Wang, Le Geng, Kai Yu, Yi Zhou, Ting Wang, Daoman Xiang, Feng Chen, et al. Automatic staging for retinopathy of prematurity with deep feature fusion and ordinal classification strategy. *IEEE Transactions on Medical Imaging*, 40(7):1750–1762, 2021.

[31] Talha Qaiser, Stefan Winzeck, Theodore Barfoot, Tara Barwick, Simon J Doran, Martin F Kaiser, Linda Wedlake, Nina Tunariu, Dow-Mu Koh, Christina Messiou, et al. Multiple instance learning with auxiliary task weighting for multiple myeloma classification. In *MICCAI*, pages 786–796. Springer, 2021.

[32] M Sadegh Saberian, Kathleen P Moriarty, Andrea D Olmstead, Christian Hallgrimson, François Jean, Ivan R Nabi, Maxwell W Libbrecht, and Ghassan Hamarneh. Deemd: Drug efficacy estimation against sars-cov-2 based on cell morphology with deep multiple instance learning. *IEEE Transactions on Medical Imaging*, 2022.

[33] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *NeurIPS*, 34, 2021.

[34] Peng Tang, Xinggang Wang, Angtian Wang, Yongluan Yan, Wenyu Liu, Junzhou Huang, and Alan Yuille. Weakly supervised region proposal network and object detection. In *ECCV*, pages 352–368, 2018.

[35] Tong Tong, Robin Wolz, Qinquan Gao, Ricardo Guerrero, Joseph V Hajnal, Daniel Rueckert, Alzheimer's Disease Neuroimaging Initiative, et al. Multiple instance learning for classification of dementia in brain mri. *Medical Image Analysis*, 18(5):808–818, 2014.

[36] Alexander Vezhnevets and Joachim M Buhmann. Towards weakly supervised semantic segmentation by means of multiple instance and multitask learning. In *CVPR*, pages 3249–3256. IEEE, 2010.

[37] Fang Wan, Chang Liu, Wei Ke, Xiangyang Ji, Jianbin Jiao, and Qixiang Ye. C-mil: Continuation multiple instance learning for weakly supervised object detection. In *CVPR*, pages 2199–2208, 2019.

[38] Xinggang Wang, Baoyuan Wang, Xiang Bai, Wenyu Liu, and Zhuowen Tu. Max-margin multiple-instance dictionary learning. In *ICML*, pages 846–854. PMLR, 2013.

[39] Xinggang Wang, Yongluan Yan, Peng Tang, Xiang Bai, and Wenyu Liu. Revisiting multiple instance neural networks. *Pattern Recognition*, 74:15–24, 2018.

[40] Yunan Wu, Arne Schmidt, Enrique Hernández-Sánchez, Rafael Molina, and Aggelos K Katsaggelos. Combining attention-based multiple instance learning and gaussian processes for ct hemorrhage detection. In *MICCAI*, pages 582–591. Springer, 2021.

[41] Gang Xu, Zhigang Song, Zhuo Sun, Calvin Ku, Zhe Yang, Cancheng Liu, Shuhao Wang, Jianpeng Ma, and Wei Xu. Camel: A weakly supervised learning framework for histopathology image segmentation. In *CVPR*, pages 10682–10691, 2019.

[42] Yan Xu, Jun-Yan Zhu, I Eric, Chao Chang, Maode Lai, and Zhuowen Tu. Weakly supervised histopathology cancer image segmentation and classification. *Medical Image Analysis*, 18(3):591–604, 2014.

[43] Dongren Yao, Jing Sui, Mingliang Wang, Erkun Yang, Yeerfan Jiaerken, Na Luo, Pew-Thian Yap, Mingxia Liu, and Dinggang Shen. A mutual multi-scale triplet graph convolutional network for classification of brain disorders using functional or structural connectivity. *IEEE Transactions on Medical Imaging*, 40(4):1279–1289, 2021.

[44] Dingwen Zhang, Deyu Meng, and Junwei Han. Co-saliency detection via a self-paced multiple-instance learning framework. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(5):865–878, 2016.

[45] Hongrun Zhang, Yanda Meng, Yitian Zhao, Yihong Qiao, Xiaoyun Yang, Sarah E Coupland, and Yalin Zheng. Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In *CVPR*, pages 18802–18812, 2022.

[46] Qi Zhang and Sally Goldman. Em-dd: An improved multiple-instance learning technique. *NeurIPS*, 14, 2001.

[47] Weijia Zhang. Non-iid multi-instance learning for predicting instance and bag labels using variational auto-encoder. *arXiv preprint arXiv:2105.01276*, 2021.

[48] Yu Zhao, Fan Yang, Yuqi Fang, Hailing Liu, Niyun Zhou, Jun Zhang, Jiarui Sun, Sen Yang, Bjoern Menze, Xinjuan Fan, et al. Predicting lymph node metastasis using histopathological images based on multiple instance learning with deep graph convolution. In *CVPR*, pages 4837–4846, 2020.

[49] Zhi-Hua Zhou, Yu-Yin Sun, and Yu-Feng Li. Multi-instance learning by treating instances as non-iid samples. In *ICML*, pages 1249–1256, 2009.