CaMIL: Causal Multiple Instance Learning for Whole Slide Image Classification

Kaitao Chen[1], **Shiliang Sun**[1*], **Jing Zhao**[1]

[1]School of Computer Science and Technology, East China Normal University, Shanghai, China
51215901053@stu.ecnu.edu.cn, shiliangsun@gmail.com, jzhao@cs.ecnu.edu.cn

Abstract

Whole slide image (WSI) classification is a crucial component in automated pathology analysis. Due to the inherent challenges of high-resolution WSIs and the absence of patch-level labels, most of the proposed methods follow the multiple instance learning (MIL) formulation. While MIL has been equipped with excellent instance feature extractors and aggregators, it is prone to learn spurious associations that undermine the performance of the model. For example, relying solely on color features may lead to erroneous diagnoses due to spurious associations between the disease and the color of patches. To address this issue, we develop a causal MIL framework for WSI classification, effectively distinguishing between causal and spurious associations. Specifically, we use the expectation of the intervention $P(Y|do(X))$ for bag prediction rather than the traditional likelihood $P(Y|X)$. By applying the front-door adjustment, the spurious association is effectively blocked, where the intervened mediator is aggregated from patch-level features. We evaluate our proposed method on two publicly available WSI datasets, Camelyon16 and TCGA-NSCLC. Our causal MIL framework shows outstanding performance and is plug-and-play, seamlessly integrating with various feature extractors and aggregators.

1 Introduction

Whole slide scanning enables the digital preservation of pathology tissues as whole slide images (WSIs), which is an essential component of modern digital pathology's automated analysis (Wang et al. 2019; Qu et al. 2022a,b). The use of deep learning models for automated WSI diagnostics holds immense potential (Qu et al. 2022c; Li, Li, and Eliceiri 2021; Campanella et al. 2019; Wang et al. 2022; Ilse, Tomczak, and Welling 2018; Lu et al. 2021; Hou et al. 2016). However, due to extremely high resolutions and a huge amount of memory, two challenges have arisen for deep learning methods. Firstly, it is infeasible to input WSIs directly into a neural network. Secondly, WSIs lack patch-level annotations because lesion areas need to be labeled by expert pathologists, which is a time-consuming process, especially when dealing with giant sizes. To address these issues, several weakly supervised methods for WSI classification have been proposed, with most methods following the multiple instance learning (MIL) formulation (Shi et al. 2020; Xiang et al. 2022; Zhu et al. 2022; Shao et al. 2021).

In the MIL paradigm, a WSI acts as the bag, while thousands of patches act as instances. A bag is considered positive if it contains at least one positive instance. The popular MIL method can be divided into three stages: 1) extracting instance-level features via a deep neural network, 2) aggregating instance-level features into a bag-level feature via a specific aggregator, and 3) training the classifier with bag labels for prediction in a fully supervised manner. Current research mainly focuses on improving aggregators and enhancing the instance-level weakly supervised signal beyond the bag-level supervision (Ilse, Tomczak, and Welling 2018; Qu et al. 2022c). Some work focuses on extracting superior features in the first stage, such as through self-supervised learning and multi-scale learning (Chen and Krishnan 2022; Chen et al. 2022; Li, Li, and Eliceiri 2021).

Currently, MIL for WSI classification has been equipped with excellent instance feature extractors and aggregators, but it is prone to learn spurious associations. The spurious association is implicit and even unobserved, but it does impair the performance of the model, which has been demonstrated in many tasks (Yang et al. 2021; Tang, Huang, and Zhang 2020; Wang et al. 2021; Ding et al. 2022). As for the WSI classification, instances in a bag exhibit high similarity because they are extracted from the same WSI and share similar tissue structures. However, significant differences exist between slides, including differences caused by imaging instruments, staining colors, or tissue types. As shown in Figure 1, the difference in color between the two samples is significant, but it is difficult to distinguish between tumors and normal cells in Figure 1b. Unfortunately, the model may make mistakes by distinguishing positive and negative bags based on superficial differences among slides (learned from the spurious association) rather than essential differences between the tumor and normal tissue (learned by eliminating the spurious association).

We formally define this spurious association as a challenge in MIL from the causal view. To explore spurious associations, we propose a causal graph based on the structural causal model (Glymour, Pearl, and Jewell 2016) to depict the problem, shown in Figure 1c. The variables we study are denoted as nodes in the graph, including instance-level feature X, bag-level feature Z, model prediction Y, and un-

[*]Corresponding Author.

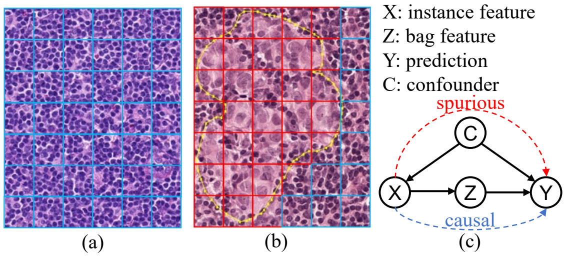


Figure 1: Samples of WSIs and the proposed causal graph explaining the MIL. (a) Negative sample. (b) Positive sample. Cancer patches are marked with a red border. Patches within the same slide are closer than the patches on different slides in color. (c) The proposed causal graph. We can identify the effect of X on Y, including the causal association (via $X \to Z \to Y$) and the confounding association (via $X \leftarrow C \to Y$). See Section 3.2 for more details.

observed confounder C. Links denote the causal relation among nodes. The link $X \to Z$ reflects that the bag-level feature is aggregated from instance-level features. The link $Z \to Y$ reflects that the classifier predicts the result from the bag-level features. In particular, the ubiquitous confounder affects the feature and prediction result (via $C \to X$ and $C \to Y$), which creates a spurious association. We ignore the image-to-feature extraction process, which is carried out in advance by a network with fixed parameters. Based on the causal graph, we can identify the effect of X on Y, including the causal association (via $X \to Z \to Y$) and the confounding association (via $X \leftarrow C \to Y$).

We noted that IBMIL (Lin et al. 2023) introduced the backdoor adjustment, where the defined observable confounded set was approximated through bag-level features, then it intervened on the bag-level features. In contrast, we model the MIL process in more detail and intervene on the instance-level features using front-door adjustment, cleverly avoiding the modeling of unobservable confounder. IBMIL adopted an independent two-stage training approach, where the confounded set was not updated during the second stage. Conversely, our training strategy is a holistic process, ensuring that all features involved are continuously updated.

In this paper, we propose a causal scheme of the multiple instance learning to face the challenge of spurious associations, named CaMIL. Our objective is to estimate the causal effect from X to Y for WSIs classification, thus we use the expectation of the intervention $P(Y|do(X))$ for bag prediction rather than the traditional likelihood $P(Y|X)$. Thanks to the causal intervention, we can block the confounding association by front-door adjustment. The key of the front-door adjustment is the mediator, and it just so happens that the bag-level feature Z acts as this critical mediator for prediction. By focusing on the mediator Z, we only condition on X in $P(Y|do(Z))$, and avoid modeling unobserved confounder C (see Section 3.2 for details). We evaluate our proposed method on two publicly available WSI datasets: Camelyon16 (Bejnordi et al. 2017) and TCGA-NSCLC. Our method is plug-and-play and compatible with various feature extractors and aggregators. Building the other SOTA methods, our method shows outstanding promise.

2 Related Work

MIL on WSIs Bag-based methods aim to train instance classifiers using pseudo-labels and subsequently aggregate the results of top-k instances to make the bag prediction (Chikontwe et al. 2020; Xu et al. 2019). In contrast to the previous methods, instance-based methods place greater emphasis on aggregating instance-level features to generate high-level bag features, which are used to train bag-level classifiers with bag labels (Sharma et al. 2021; Zhu et al. 2017; Li, Li, and Eliceiri 2021; Zhang et al. 2022; Ilse, Tomczak, and Welling 2018; Lu et al. 2021; Shao et al. 2021). These methods are empirically proven to be superior to instance-based methods. A classical method is ABMIL (Ilse, Tomczak, and Welling 2018), which uses attention scores as weight values to aggregate individual instance-level features. Recently, researchers have explored modifications to the weight generation, such as using the cosine distance between instances and the target instance (Li, Li, and Eliceiri 2021). Moreover, TransMIL (Shao et al. 2021) relies on correlated multiple instance learning via transformers (Vaswani et al. 2017) to capture interactive information within a WSI. Another strategy, introduced by Lu *et al.* (Lu et al. 2021), incorporates a clustering loss to expand the distance between positive and negative instances within each bag. Nevertheless, despite these advances, spurious associations are almost ignored in WSIs and instance-based methods may suffer from spurious associations, which could limit their efficacy. Consequently, a new causal method is necessary to address this challenge and improve the performance of WSI classification.

Causal Inference Combining machine learning with causal inference is relatively new research that has gained significant attention in recent years. One key advantage of causal inference is the ability to eliminate harmful confounding effects through interventions using structural causal models (Glymour, Pearl, and Jewell 2016; Pearl 2014). Several studies have incorporated causal inference into various machine learning tasks, such as semantic segmentation (Zhang et al. 2020), visual categorization (Rao et al. 2021), vision-language tasks (Yang et al. 2021; Niu et al. 2021; Wang et al. 2020), and reinforcement learning (Everitt et al. 2021; Pearl 2014). We noted that IBMIL (Lin et al. 2023) used the backdoor adjustment and intervened on bags, yet it obtained the representation of the confounded set through bag-level features. In contrast, our approach employs front-door adjustment by intervening on instances and avoids the necessity of directly modeling the confounded set (see Section 4.5 for details).

3 Method

3.1 Multiple Instance Learning Formulation

In multiple instance learning, each bag consists of multiple instances. The bag is labeled negative only if all instances in the bag are negative. Conversely, the bag is labeled positive if at least one of its instances is positive. Let

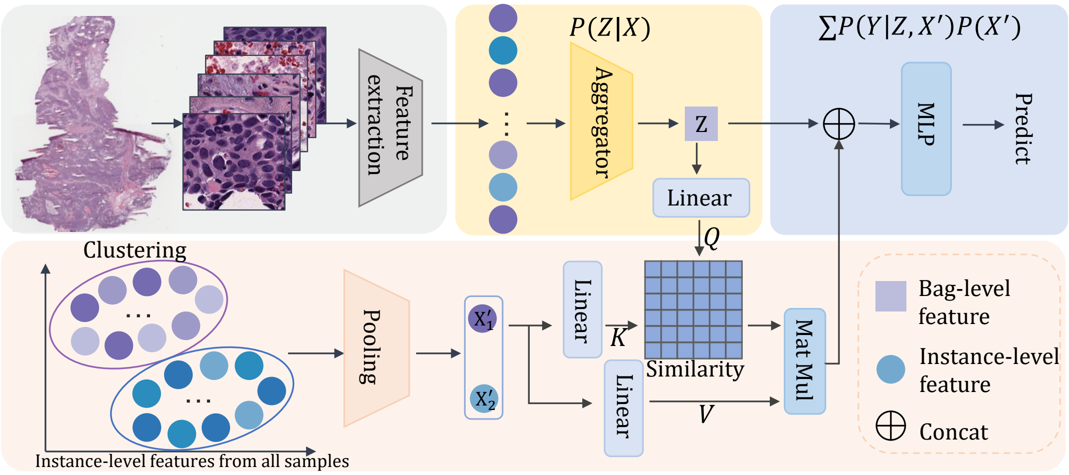


Figure 2: Overall architecture of our proposed causal multiple instance learning (CaMIL). The top part represents conventional MIL, and the lower part illustrates the key of CaMIL. Instances within all bags are grouped into clusters, and then the pooling features of each cluster are obtained. Subsequently, a cross-attention mechanism is employed to fuse these features with the variable Z, enabling better feature representation and causal modeling for improved bag-level predictions. The instance-level features used for clustering are continuously updated.

$S_i = \{(p_{i,1}, y_{i,1}), ..., (p_{i,n}, y_{i,n})\}$ represent a bag with $p_{i,j}$ instances and their corresponding labels $y_{i,j} \in \{0, 1\}$. The bag label Y_i can be defined as follows (Ilse, Tomczak, and Welling 2018):

$$Y_i = \begin{cases} 0, & \text{iff } \sum_j y_{i,j} = 0, \\ 1, & \text{otherwise.} \end{cases} \tag{1}$$

We only have access to bag-level labels, and the instance-level labels remain unknown. The objective is to learn a model that can predict the label of the bag based on the input instances. As shown in Figure 2, the conventional MIL mainly consists of three modules: extraction, aggregation, and classification, corresponding to the top part in the figure. Bag-based multiple instance learning methods employ an instance-level feature extractor $f(\cdot)$ that maps the instance into the low-dimensional feature representation. An aggregation $g(\cdot)$ is then applied to combine these instance-level features and to produce a bag-level feature. Finally, a classifier $h(\cdot)$ predicts the probabilities $\hat{Y}$ of the bag label based on the bag-level feature. The whole stage can be expressed by the following formula:

$$\hat{Y}_i = h\left(g(x_{i,1}, x_{i,n}, ..., x_{i,n})\right), x_{i,j} = f(p_{i,j}). \tag{2}$$

The aggregator and classifier are eventually optimized by minimizing the cross-entropy between Y_i and $\hat{Y}_i$.

3.2 CaMIL

To investigate the multiple instance learning methods and what factors affect the prediction, we propose a causal graph with variables: instance-level feature X, bag-level feature Z, model prediction Y, and unobserved confounder C, shown in Figure 3. The causal graph is used to portray the study variables of interest and their interactions through causal links. The link $X \rightarrow Z$ reflects that the bag-level feature is aggregated from instance-level features, and the link $Z \rightarrow Y$ reflects that the classifier predicts the result from the bag-level features. In particular, the ubiquitous confounder affects the feature and prediction result (via $C \rightarrow X$ and $C \rightarrow Y$). Note that we ignore the image-to-feature extraction process, which is carried out in advance by a network with fixed parameters. We can intuitively identify the effect of X on Y in the graph, including the causal association (via $X \rightarrow Z \rightarrow Y$) and the confounding association (via $X \leftarrow C \rightarrow Y$). The confounding association can be a harmful shortcut for models, as models can make predictions directly without exploring whether positive instances exist in the input data. In the causal association, the mediator Z is very helpful, and we can split the causal effect into three steps by focusing on Z (Glymour, Pearl, and Jewell 2016; Pearl 2014): 1) calculate the causal effect of $X \rightarrow Z$, 2) calculate the causal effect of $Z \rightarrow Y$, and 3) calculate the causal effect of $X \rightarrow Z \rightarrow Y$ by combining the above steps.

Step I. As shown in Figure 3a, it is easy to calculate the causal effect of X on Z, because the path $X \leftarrow C \rightarrow Y \leftarrow Z$ is a blocked back-door path and Y is a collider. The only causal effect flows via $X \rightarrow Z$:

$$P(z \mid do(x)) = P(z \mid x). \tag{3}$$

Step II. It is critical to calculate the causal effect of Z on Y shown in Figure 3b, including the causal association (via $Z \rightarrow Y$) and the confounding association (via $Z \leftarrow X \leftarrow C \rightarrow Y$). By Bayesian network factorization and conditional independence, we can transform the unsolvable C into X. Importantly, we condition on X and marginalize it out to block the confounding association:

$$P(y \mid do(z)) = \sum_x P(y \mid z,\ x)P(x). \tag{4}$$

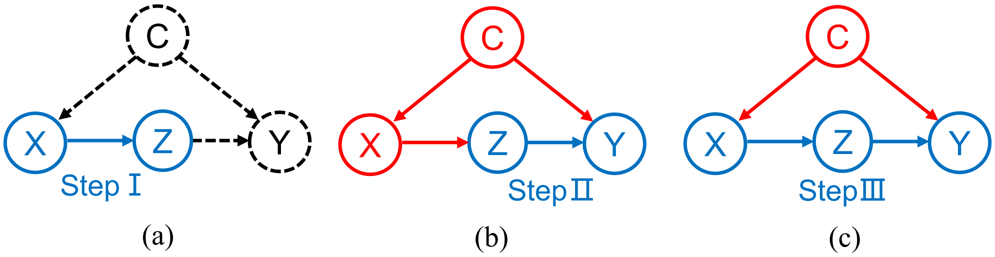


Figure 3: The proposed causal graph explaining the multiple instance learning. (a) The causal effect of $X \rightarrow Z$. The dotted link $X \leftarrow C \rightarrow Y \leftarrow Z$ denotes a blocked path because of the collider Y. (b) The causal effect of $Z \rightarrow Y$. The red link $Z \leftarrow X \leftarrow C \rightarrow Y$ denotes a back-door path. (c) The causal effect of $X \rightarrow Y$.

Step III. The causal effect of X on Z and the causal effect of Z on Y are obtained, then we can combine the two steps to get the total causal effect shown in Figure 3c. We sum over all possible realizations of the random variable Z:

$$P(y \mid do(x)) = \sum_{z} P(z \mid x) \sum_{x'} P(y \mid z,\, x')P(x'). \quad (5)$$

3.3 Framework

Instances within the bag are encoded into instance-level features $\{x_{i,1},\, x_{i,2},\, ...,\, x_{i,n}\}$. As there are tens of thousands of instances in the bag, this step is done in advance. An aggregation $g(\cdot)$ combines these instance-level features into a bag-level feature z, which corresponds to the term $P(z \mid x)$ in Equation (5). We keep in mind that the variable z depends on x by an aggregation $g(\cdot)$, so $z = g(x)$ is omitted as z in the subsequent calculation:

$$\begin{aligned} P(y \mid do(x)) &= \sum_{x'} P\left(y \mid z = g(x),\, x'\right) \cdot P(x') \\ &= \mathbb{E}_{x'}\left[P\left(y \mid z,\, x'\right)\right]. \end{aligned} \quad (6)$$

Now, solving $\sum_{x'} P(y \mid z,\, x')P(x')$ is our goal. We introduce an instance buffer to store instances within all bags in advance. However, it is computationally expensive for each bag prediction because of the iteration over all instances. Instead, it is more reasonable to first divide all instances within the buffer into k clusters using K-means and then iterate through each cluster. Each cluster represents a subset of instances that share similar characteristics, thereby capturing the essential information needed for bag classification. Then, the average features of each cluster are obtained through pooling operations. Therefore, we obtain the new set $x' = [x'_1,\, x'_2,\, ...,\, x'_K]$ of instance-level features after clustering and pooling, where x'_k is the average of the instance-level features in k-th cluster. The final computational complexity is reduced from $O(N \cdot N \cdot M)$ to $O(N \cdot K)$, where N is the number of bags, M is the average number of instances in each bag, and K is the number of clusters. Subsequently, a cross-attention mechanism (Vaswani et al. 2017) is employed to fuse the average features of each cluster with the variable Z. This fusion allows for the integration of global information, and the formula of cross-attention is as follows:

$$\phi(z,\, x') = \left[P(x') \cdot \mathrm{Softmax}\left(\frac{(W_q z)^{\top}(W_k x')}{\sqrt{d}}\right)\right](W_v x'), \quad (7)$$

where W_q, W_k, and W_v are three linear projections, and d denotes the dimension of feature. Generally, we set the prior $P(x')$ to $1/K$, which is fair for each cluster. The concat operator is used to merge the bag-level feature and the fused feature, and then we apply NWGM (Xu et al. 2015; Baldi and Sadowski 2014) to approximate the expectation and to circumvent multiple forward passes within the causal network. The formula is as follows:

$$P(y \mid do(x)) \approx P\left(y \mid z \oplus \phi(z,\, x')\right). \quad (8)$$

The feature extractor $f(\cdot)$ is a neural network using fixed parameters, including ResNet (He et al. 2016) or ViT (Dosovitskiy et al. 2020). The aggregation $g(\cdot)$ is a permutation-invariant operator (Ilse, Tomczak, and Welling 2018) that corresponds to the attention-based method and other variants. The classifier $h(\cdot)$ is a multi-layer perception.

Note that IBMIL (Lin et al. 2023) fused the bag-level feature and confounded set obtained by clustering the bag-level features. Instead, we combine the bag-level feature z and global instance-level features x' by Equation (7). More importantly, the aggregation of z is conducted by $g(\cdot)$, and we demonstrate that it is more advantageous to sample z multiple times with distinct aggregators $g(\cdot)$ in Section 4.5.

4 Experiments and Results

4.1 Datasets and Evaluation Metrics

Camelyon16 (Bejnordi et al. 2017) is a public dataset for the detection of metastases in breast cancer. The dataset consists of 270 training images and 129 test images, which are cropped to roughly 1.2 million non-overlapping patches of size 256 × 256 at 5× magnification, with an average of approximately 3,000 patches obtained per WSI.

The cancer genome atlas non-small cell lung cancer (**TCGA-NSCLC**) dataset includes two subtypes of lung cancer, lung squamous cell carcinoma (LUSC) and lung adenocarcinoma (LUAD), with a total of 1,054 WSI. We divide the dataset into a training set, a validation set, and a test set according to 7:1:2. The dataset yields 14.9 million patches at 20× magnification, with an average of roughly 14,000 patches obtained per bag.

To measure our method more comprehensively, we use four evaluation metrics of classification performance, including precision, accuracy, F1, and area under the curve

	ResNet18				ViT-small			
	Prec.	Acc.	F1	AUC	Prec.	Acc.	F1	AUC
ABMIL	81.71	80.62	79.94	80.96	77.75	77.52	77.59	80.81
w/ causal	83.43	82.17	81.54	82.47	80.08	79.84	79.45	80.89
Δ	1.72	1.55	1.60	1.51	2.33	2.32	1.86	0.08
DSMIL	78.97	79.07	78.98	81.08	81.51	81.40	81.12	85.55
w/ causal	85.41	82.17	81.11	82.37	88.65	87.60	87.28	88.14
Δ	6.44	3.10	2.13	1.29	7.14	6.20	6.16	2.59
TransMIL	69.44	69.77	69.39	69.15	72.26	72.09	72.16	76.50
w/ causal	72.26	72.09	72.16	72.60	75.03	75.19	75.04	76.23
Δ	2.82	2.32	2.77	3.45	2.77	3.10	2.88	-0.27
CLAM-SB	83.92	79.84	78.36	77.16	85.17	82.95	82.15	82.69
w/ causal	83.38	81.40	80.53	81.85	85.73	83.72	83.03	82.67
Δ	-0.54	1.56	2.17	4.69	0.56	0.77	0.88	-0.02
CLAM-MB	82.23	79.84	78.74	77.62	84.07	81.40	80.38	82.81
w/ causal	85.17	82.95	82.15	81.77	86.43	83.72	82.90	85.53
Δ	2.94	3.11	3.41	4.15	2.36	2.32	2.52	2.72
DTFD	81.02	80.62	80.18	78.79	82.10	82.17	82.09	81.35
w/ causal	87.45	86.05	85.60	85.63	87.28	83.72	82.76	82.32
Δ	6.43	5.43	5.42	6.84	5.18	1.55	0.67	0.97

Table 1: Results on the Camelyon16 dataset (%).

(AUC). The setting of the classification threshold is the same as in DSMIL (Li, Li, and Eliceiri 2021).

4.2 Implementation Details

For the pre-processing, two extractors of different architectures are used, including a CNN-based ResNet18 (He et al. 2016) and a transformer-based ViT-small (Dosovitskiy et al. 2020). Both ResNet18 and ViT-small use pre-trained parameters. The difference is that ResNet18 extracts a 512-dimensional feature, while ViT-small extracts multiple tokens in 384 dimensions, with only the class token being utilized. During training, instance-level features are mapped to 512-dimensional space via a fully connected layer. We use the Adam optimizer with the initial learning rate of 2e-4 and weight decay of 5e-4 for the majority of experiments. In particular, we remain the Lookahead optimizer (Zhang et al. 2019) in the experiments about TransMIL (Shao et al. 2021), consistent with the setup described in the original paper. The size of the mini-batch for training models is 1, and the number of epochs is 100. By default, up to 4,000 instances are randomly selected from each bag and added to the instance buffer, with the buffer being updated every epoch. The number of clusters by K-means is set to 8. All experiments are implemented on an NVIDIA GeForce RTX 3090.

4.3 Experimental Results

We integrate our causal method into the SOTA MIL models (Ilse, Tomczak, and Welling 2018; Li, Li, and Eliceiri 2021; Shao et al. 2021; Lu et al. 2021; Zhang et al. 2022) on the task of WSI classification. Table 1 shows the results on the Camelyon16 dataset for WSI classification. Overall, the performance of the original methods has been consistently enhanced after the fusion of the causal method on the metrics of precision, accuracy, F1, and AUC. Apparently, the performance improvement of four metrics under ResNet18 extraction exceeds 5% on the DTFD method. Importantly, spurious association effects are universal in different architectures, and causal methods work in both a CNN-based ResNet18 (He et al. 2016) and a transformer-based ViT-small (Dosovitskiy et al. 2020).

Table 2 shows the results of the TCGA-NSCLC dataset. Similar to the findings from the Camelyon16 dataset, the performance of experiments has been improved, which proves that our method is compatible with various feature extractors and aggregators. Specifically, the DTFD method exhibits a more significant improvement compared to other MIL aggregators when using the ResNet18 extractor, consistent with the results obtained from the Camelyon16 dataset. Under the different MIL aggregators, the causal method has a more significant improvement on the ResNet18 extractor than on the ViT-small extractor.

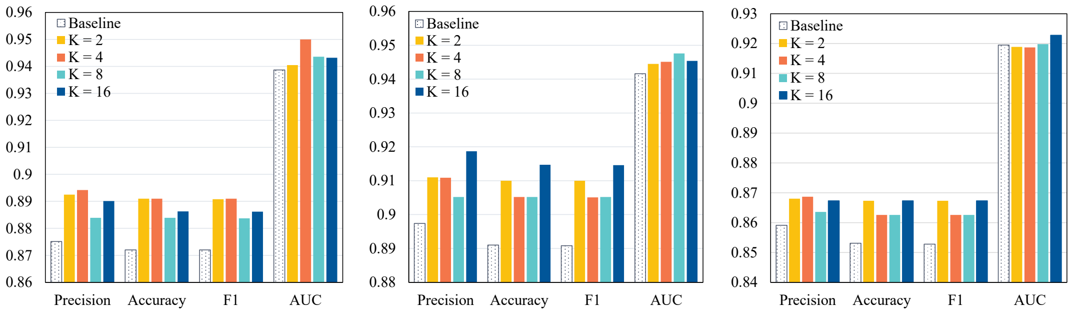
4.4 Ablation Study and Sensitivity Analysis

Number of Clusters. The number of clusters is a hyperparameter, and we respectively set the number of clusters to 2, 4, 8, or 16. We study the effect of the number of clusters based on the ABMIL method on the TCGA-NSCLC dataset. Our method remains effective even when the number of clusters reaches 16. Moreover, we also carry out this experiment on two non-parametric aggregators: max-pooling and mean-pooling. Figure 4 shows that our method improves the results under different numbers of clusters, regardless of using a parameterized aggregator or a non-parametric aggregator.

Number of Instances for Clustering. We explore the effect of the number of instances in each bag that enters the instance buffer on the TCGA-NSCLC dataset. Figure 5a

	ResNet18				ViT-small			
	Prec.	Acc.	F1	AUC	Prec.	Acc.	F1	AUC
ABMIL	85.47	85.67	85.51	90.94	87.51	87.20	87.20	93.87
w/ causal	87.93	87.92	87.92	91.82	88.39	88.39	88.37	94.36
Δ	2.46	2.25	2.41	0.88	0.88	1.19	1.17	0.49
DSMIL	76.61	76.54	76.54	81.38	84.57	84.12	84.08	89.41
w/ causal	78.33	78.20	78.17	82.60	84.50	84.36	84.33	89.61
Δ	1.72	1.66	1.63	1.22	-0.07	0.24	0.25	0.20
TransMIL	86.87	86.73	86.73	92.25	88.40	88.15	88.11	92.41
w/ causal	88.46	88.15	88.15	91.70	90.07	90.05	90.04	94.59
Δ	1.59	1.42	1.42	-0.55	1.67	1.90	1.93	2.18
CLAM-SB	83.55	81.99	81.86	88.79	88.79	88.15	88.07	94.67
w/ causal	86.91	86.73	86.69	91.62	90.46	90.05	90.00	95.64
Δ	3.36	4.74	4.83	2.83	1.67	1.90	1.93	0.97
CLAM-MB	84.98	83.89	83.81	89.08	86.70	86.26	86.24	94.36
w/ causal	87.25	87.20	87.21	91.18	89.58	88.63	88.59	95.02
Δ	2.27	3.31	3.40	2.10	2.88	2.37	2.35	0.66
DTFD	83.43	82.94	82.92	88.92	87.40	87.20	87.20	90.52
w/ causal	87.92	87.92	87.92	91.47	90.05	90.05	90.05	93.20
Δ	4.49	4.98	5.00	2.55	2.65	2.85	2.85	2.68

Table 2: Results on the TCGA-NSCLC dataset (%).



(a) (b) (c)

Figure 4: Results of the ablation. (a) Number of clusters on the ABMIL method. (b) Number of clusters on the max-pooling aggregators. (c) Number of clusters on the mean-pooling aggregators.

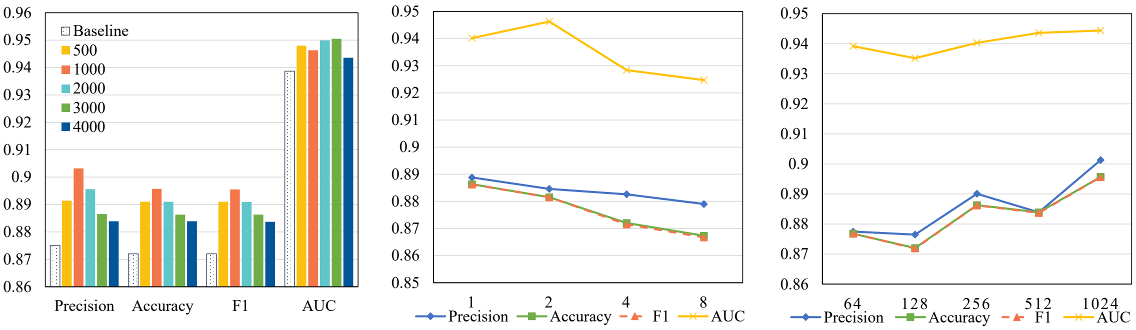
shows the results based on the ABMIL method. We find that we only need to sample 1,000 instances of each bag into the instance buffer to get good results when the average bag actually has 14,000 instances.

Frequency of Update. To verify the effect of the update frequency of the instance buffer, we respectively set the instance buffer to be updated every 1, 2, 4, or 8 epochs. Figure 5b shows the results based on the ABMIL method on the TCGA-NSCLC dataset. When the instance buffer is updated slowly, the performance of the model drops because the features in the instance buffer still remain in the same state as before several epochs.

Dimension of Instance-level Features. All the instance features are initially mapped to the task-specific feature through a linear layer, and the dimension of the task-specific feature affects the number of parameters in the model. We respectively set dimensions to 64, 128, 256, 512, and 1,024 based on the TCGA-NSCLC dataset. Changes in precision and accuracy remain consistent in Figure 5c, and the best performance is achieved when the dimension rises to 1,024.

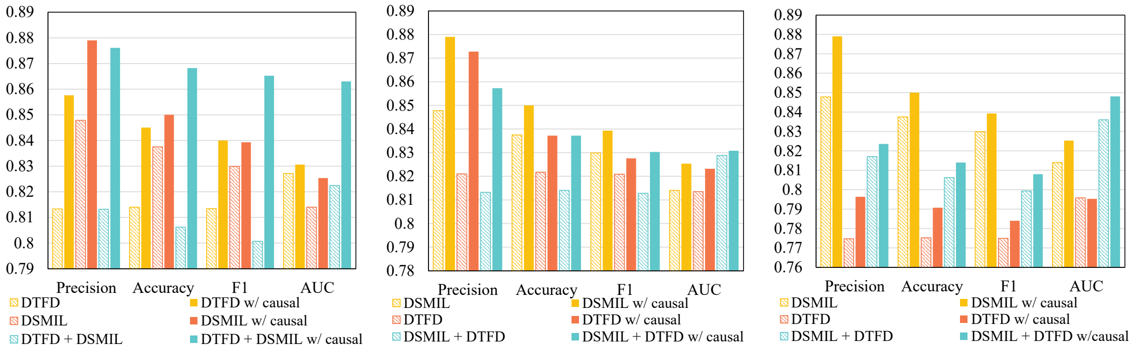
4.5 Discussion

While IBMIL (Lin et al. 2023) leverages backdoor adjustment and acquires the representation of the confounded set through clustering bag-level features, our approach employs front-door adjustment, obviating the necessity of directly modeling the confounder set. In our methodology, we perform clustering on instance-level features, aiming to decrease the instance count for intervention expectation calculation. IBMIL employs a two-stage training paradigm. In the second stage, the deconfounding network is trained, with the confounded set remaining constant during this phase. In con-



(a) (b) (c)

Figure 5: Sensitivity analysis. (a) Number of instances for clustering. (b) Frequency of update. (c) Dimension of features.



(a) (b) (c)

Figure 6: Results of fusing multiple aggregators. (a) The combination of DSMIL and DTFD-MaxS (without or with the causal method). (b) The combination of DSMIL and DTFD-MaxMinS. (c) The combination of DSMIL and DTFD-AFS.

trast, our approach is single-stage. The instances needed for clustering evolve with network updates, and the clustering outcomes are renewed with each epoch. We provide the experiments with IBMIL based on DSMIL, and Table 3 shows that the enhancement of our method is pronounced.

	Prec.	Acc.	F1	AUC
IBMIL	81.99	81.40	80.92	83.85
CaMIL	85.41	82.17	81.11	82.37

Table 3: Comparison of the results of CaMIL and IBMIL.

In the above experiments, the bag-level features are obtained through an aggregator. Ensemble learning assembles multiple basic learning models into a strong learner. To explore whether ensemble aggregators can mitigate the effect of spurious associations, we integrate several different aggregators and then use the average of the prediction results of multiple classifiers as the final prediction. We integrate DSMIL and three different strategies of DTFD separately and conduct experiments on the Camelyon16 dataset. As shown in Figure 6, in the absence of causal intervention, using different aggregators may not exceed the best experimental performance achieved by a single aggregator, and may even yield worse results than those of all individual aggregators. This phenomenon may be due to the presence of similarities or common biases among the aggregators, which are magnified. After incorporating causal methods, the aforementioned unfavorable phenomenon is significantly improved, especially in the combinations of DTFD-MaxS and DSMIL, which greatly surpasses the best experimental performance achieved by a single aggregator.

5 Conclusion

In this paper, we establish a causal MIL framework to face the challenge of spurious associations. We not only theoretically explain the causal effect by distinguishing between causal and confounding associations, but also derive a causal intervention for unobserved confounders. By focusing on the bag-level feature and acting as the mediator, we use the front-door adjustment to block the spurious association and calculate the expectation of interventions on the mediator for bag predictions instead of the traditional likelihood. Extensive experiments are conducted on two publicly available WSI datasets: Camelyon16 and TCGA-NSCLC. Our causal method is a plug-and-play framework and it is applicable to different feature extractors and aggregators. Building the other SOTA methods, CaMIL shows outstanding promise.

Acknowledgments

This work was supported by NSFC Projects 62076096 and 62006078, the STCSM Project 22ZR1421700, Shanghai Knowledge Service Platform Project ZF1213, the Open Research Fund of KLATASDS-MOE and the Fundamental Research Funds for the Central Universities.

References

Baldi, P.; and Sadowski, P. 2014. The dropout learning algorithm. *Artificial Intelligence*, 210: 78–122.

Bejnordi, B. E.; Veta, M.; Van Diest, P. J.; Van Ginneken, B.; Karssemeijer, N.; Litjens, G.; Van Der Laak, J. A.; Hermsen, M.; Manson, Q. F.; Balkenhol, M.; et al. 2017. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22): 2199–2210.

Campanella, G.; Hanna, M. G.; Geneslaw, L.; Miraflor, A.; Werneck Krauss Silva, V.; Busam, K. J.; Brogi, E.; Reuter, V. E.; Klimstra, D. S.; and Fuchs, T. J. 2019. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 25(8): 1301–1309.

Chen, R. J.; Chen, C.; Li, Y.; Chen, T. Y.; Trister, A. D.; Krishnan, R. G.; and Mahmood, F. 2022. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16144–16155.

Chen, R. J.; and Krishnan, R. G. 2022. Self-supervised vision transformers learn visual concepts in histopathology. *arXiv preprint arXiv:2203.00585*.

Chikontwe, P.; Kim, M.; Nam, S. J.; Go, H.; and Park, S. H. 2020. Multiple instance learning with center embeddings for histopathology classification. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, 519–528.

Ding, L.; Yu, D.; Xie, J.; Guo, W.; Hu, S.; Liu, M.; Kong, L.; Dai, H.; Bao, Y.; and Jiang, B. 2022. Word embeddings via causal inference: Gender bias reducing and semantic information preserving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 11864–11872.

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

Everitt, T.; Hutter, M.; Kumar, R.; and Krakovna, V. 2021. Reward tampering problems and solutions in reinforcement learning: A causal influence diagram perspective. *Synthese*, 198(27): 6435–6467.

Glymour, M.; Pearl, J.; and Jewell, N. P. 2016. *Causal inference in statistics: A primer*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 770–778.

Hou, L.; Samaras, D.; Kurc, T. M.; Gao, Y.; Davis, J. E.; and Saltz, J. H. 2016. Patch-based convolutional neural network for whole slide tissue image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2424–2433.

Ilse, M.; Tomczak, J.; and Welling, M. 2018. Attention-based deep multiple instance learning. In *International Conference on Machine Learning*, 2127–2136.

Li, B.; Li, Y.; and Eliceiri, K. W. 2021. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14318–14328.

Lin, T.; Yu, Z.; Hu, H.; Xu, Y.; and Chen, C.-W. 2023. Interventional bag multi-instance learning on whole-slide pathological images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Lu, M. Y.; Williamson, D. F.; Chen, T. Y.; Chen, R. J.; Barbieri, M.; and Mahmood, F. 2021. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5(6): 555–570.

Niu, Y.; Tang, K.; Zhang, H.; Lu, Z.; Hua, X.-S.; and Wen, J.-R. 2021. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12700–12710.

Pearl, J. 2014. Interpretation and identification of causal mediation. *Psychological Methods*, 19: 459.

Qu, L.; Liu, S.; Liu, X.; Wang, M.; and Song, Z. 2022a. Towards label-efficient automatic diagnosis and analysis: a comprehensive survey of advanced deep learning-based weakly-supervised, semi-supervised and self-supervised techniques in histopathological image analysis. *Physics in Medicine & Biology*.

Qu, L.; Luo, X.; Liu, S.; Wang, M.; and Song, Z. 2022b. Dgmil: Distribution guided multiple instance learning for whole slide image classification. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, 24–34.

Qu, L.; Wang, M.; Song, Z.; et al. 2022c. Bi-directional weakly supervised knowledge distillation for whole slide image classification. *Advances in Neural Information Processing Systems*, 35: 15368–15381.

Rao, Y.; Chen, G.; Lu, J.; and Zhou, J. 2021. Counterfactual attention learning for fine-grained visual categorization and re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1025–1034.

Shao, Z.; Bian, H.; Chen, Y.; Wang, Y.; Zhang, J.; Ji, X.; et al. 2021. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in Neural Information Processing Systems*, 34: 2136–2147.

Sharma, Y.; Shrivastava, A.; Ehsan, L.; Moskaluk, C. A.; Syed, S.; and Brown, D. 2021. Cluster-to-conquer: A framework for end-to-end multi-instance learning for whole slide image classification. In *Medical Imaging with Deep Learning*, 682–698.

Shi, X.; Xing, F.; Xie, Y.; Zhang, Z.; Cui, L.; and Yang, L. 2020. Loss-based attention for deep multiple instance learning. In *Proceedings of the AAAI conference on Artificial Intelligence*, volume 34, 5742–5749.

Tang, K.; Huang, J.; and Zhang, H. 2020. Long-tailed classification by keeping the good and removing the bad momentum causal effect. *Advances in Neural Information Processing Systems*, 33: 1513–1524.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Wang, T.; Huang, J.; Zhang, H.; and Sun, Q. 2020. Visual commonsense R-CNN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10760–10770.

Wang, T.; Zhou, C.; Sun, Q.; and Zhang, H. 2021. Causal attention for unbiased visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3091–3100.

Wang, X.; Chen, H.; Gan, C.; Lin, H.; Dou, Q.; Tsougenis, E.; Huang, Q.; Cai, M.; and Heng, P.-A. 2019. Weakly supervised deep learning for whole slide lung cancer image analysis. *IEEE Transactions on Cybernetics*, 50(9): 3950–3962.

Wang, X.; Xiang, J.; Zhang, J.; Yang, S.; Yang, Z.; Wang, M.-H.; Zhang, J.; Yang, W.; Huang, J.; and Han, X. 2022. SCL-WC: Cross-slide contrastive learning for weakly-supervised whole-slide image classification. *Advances in Neural Information Processing Systems*, 35: 18009–18021.

Xiang, T.; Song, Y.; Zhang, C.; Liu, D.; Chen, M.; Zhang, F.; Huang, H.; O'Donnell, L.; and Cai, W. 2022. Dsnet: A dual-stream framework for weakly-supervised gigapixel pathology image analysis. *IEEE Transactions on Medical Imaging*, 41(8): 2180–2190.

Xu, G.; Song, Z.; Sun, Z.; Ku, C.; Yang, Z.; Liu, C.; Wang, S.; Ma, J.; and Xu, W. 2019. Camel: A weakly supervised learning framework for histopathology image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10682–10691.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, 2048–2057.

Yang, X.; Zhang, H.; Qi, G.; and Cai, J. 2021. Causal attention for vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9847–9857.

Zhang, D.; Zhang, H.; Tang, J.; Hua, X.-S.; and Sun, Q. 2020. Causal intervention for weakly-supervised semantic segmentation. *Advances in Neural Information Processing Systems*, 33: 655–666.

Zhang, H.; Meng, Y.; Zhao, Y.; Qiao, Y.; Yang, X.; Coupland, S. E.; and Zheng, Y. 2022. Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18802–18812.

Zhang, M.; Lucas, J.; Ba, J.; and Hinton, G. E. 2019. Lookahead optimizer: k steps forward, 1 step back. *Advances in Neural Information Processing Systems*, 32.

Zhu, W.; Lou, Q.; Vang, Y. S.; and Xie, X. 2017. Deep multi-instance networks with sparse label assignment for whole mammogram classification. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, 603–611.

Zhu, Z.; Yu, L.; Wu, W.; Yu, R.; Zhang, D.; and Wang, L. 2022. MuRCL: Multi-instance Reinforcement Contrastive Learning for Whole Slide Image Classification. *IEEE Transactions on Medical Imaging*.