Interventional Multi-Instance Learning with Deconfounded Instance-Level Prediction

Tiancheng Lin\textsuperscript{1,2}, Hongteng Xu\textsuperscript{3,4,5}, Canqian Yang\textsuperscript{1,2} and Yi Xu\textsuperscript{1,2*}

\textsuperscript{1} Shanghai Key Lab of Digital Media Processing and Transmission, Shanghai Jiao Tong University
\textsuperscript{2} MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
\textsuperscript{3} Gaoling School of Artificial Intelligence, Renmin University of China
\textsuperscript{4} Beijing Key Laboratory of Big Data Management and Analysis Methods
\textsuperscript{5} JD Explore Academy
\{ltc19940819, charles.young, xuyi\}@sjtu.edu.cn, hongtengxu@ruc.edu.cn

\section*{Abstract}
When applying multi-instance learning (MIL) to make predictions for bags of instances, the prediction accuracy of an instance often depends on not only the instance itself but also its context in the corresponding bag. From the viewpoint of causal inference, such bag contextual prior works as a confounder and may result in model robustness and interpretability issues. Focusing on this problem, we propose a novel interventional multi-instance learning (IMIL) framework to achieve deconfounded instance-level prediction. Unlike traditional likelihood-based strategies, we design an Expectation-Maximization (EM) algorithm based on causal intervention, providing a robust instance selection in the training phase and suppressing the bias caused by the bag contextual prior. Experiments on pathological image analysis demonstrate that our IMIL method substantially reduces false positives and outperforms state-of-the-art MIL methods.

\section*{Introduction}
In many real-world scenarios, fine-grained labels of data, e.g., pixel-wise annotations of high-resolution images, are often unavailable due to the limitations of human resources, time, and budgets. To mitigate the requirement for high-quality labels, multi-instance learning (MIL) treats multiple instances as a bag and learns an instance-level predictive model from a set of labeled bags (Dietterich, Lathrop, and Lozano-Pérez 1997). Such a paradigm has been widely used in many applications, e.g., image classification (Wu et al. 2015a), object detection (Wan et al. 2019), semantic segmentation (Xu et al. 2019), etc. Among them, whole slide pathological image (WSI) classification is a representative example. Each WSI is a bag with a pathological label, and the patches of the WSI are unlabeled instances in the bag. The MIL framework learns an instance-level classifier to indicate the patches corresponding to the lesions.

Although many MIL methods have been proposed and achieved encouraging performance in extensive applications (Hou et al. 2016; Chen et al. 2019; Chikontwe et al. 2020), they often suffer from an issue called “bag contextual prior”. In particular, the bag contextual prior is a kind of instance-shared information corresponding to bags but irrelevant to their instances, which may be inherited by models like deep neural networks (DNNs) and lead to questionable instance-level prediction. Figure 1(a) illustrates the bag contextual prior in WSI classification. In each bag, its patches (instances) with different labels (positive/negative) often have similar attributes on color and texture. For different bags, on the contrary, their patches with the same labels can be very different in vision. As a result, the “similarity” within each bag and the “difference” across different bags, which harm instance-level prediction, could be wrongly exploited by MIL models. As shown in Figure 1(b), the models predict similar scores for the instances in the same bag. From the viewpoint of causal inference, the above bag contextual prior is a confounder that causes the spurious correlation between instances and labels, making the prediction depend on both the key instance and its useless context. Therefore, a robust and interpretable MIL model should build an efficient mechanism to suppress the bias caused by the bag contextual prior, predicting the classification scores via revealing the actual causality between instances and labels.

To achieve deconfounded instance-level prediction, we propose a novel interventional multi-instance learning (IMIL), in which a structural causal model (SCM) (Pearl, Glymour, and Jewell 2016) analyzes the causalities among the bag contextual prior, instances, and labels. As depicted in Figure 2, our IMIL is an Expectation-Maximization (EM) algorithmic framework with two novel strategies for con-
founder bias removal and robust instance selection. In the training phase, we initially assign the instances with the label of the bag they belong to, then we alternate the following E-step and M-step until convergence. In the M-step, we apply physical interventions to remove the confounder bias, where various data augmentations are adopted. In the E-step, we approximate the total effect of our model, where we first reweight the scores of instances to get the de-biased prediction and then select instances via both direct causal effect and indirect mediator effect. Note that, different from the existing instance selection criteria (Hou et al. 2016; Chen et al. 2019; Wang et al. 2019a), our model approximates the causal effect without external information.

We compare our IMIL framework with state-of-the-art MIL methods through the lens of causality and analyze their connections and differences in detail. The effectiveness of our IMIL is verified over two public WSI datasets, i.e., DigestPath (Li et al. 2019a) and Camelyon16 (Bejnordi et al. 2017). Experimental results show that our IMIL achieves superior performance in the WSI classification tasks. Significantly, the proposed physical intervention is compatible with all compared MIL methods, bringing consistent performance boosting. Furthermore, we demonstrate the potentials of IMIL to multi-class multi-label MIL problems on the Pascal VOC dataset (Everingham et al. 2015).

Related Work

Bag-level MIL

Bag-level MIL either implicitly utilizes bag-to-bag distance/similarity or explicitly trains a bag classifier (Wang et al. 2018), whose novel distance metrics and aggregation operators are designed based on various neural network architectures (Nazeri, Aminpour, and Ebrahimi 2018; Wang et al. 2019b; Zhao et al. 2020a), pooling strategies (Yan et al. 2018), and attention mechanisms (Ilse, Tomczak, and Welling 2018; Shi et al. 2020). For large-scale MIL scenarios like gigapixel image analysis, we often have to implement the bag-level MIL models by a two-stage strategy (Campanella et al. 2019; Tellez et al. 2019; Li et al. 2019b; Zhao et al. 2020b; Yao et al. 2020; Li, Li, and Eliceiri 2021), training an instance-level feature extractor, and then aggregating instance features as bag-level representations.

Instance-level MIL

Instance-level MIL is a natural solution to gigapixel image analysis, where a classifier is trained to produce a score for each instance, and the instance scores are aggregated to produce a bag score. The representative method is SimpleMIL, which directly propagates the bag label to its instances (Ray and Craven 2005; Cheplygina et al. 2017). To suppress the noise caused by the instance-level supervision, the work in (Wang et al. 2019a) directly introduces extra cleaner annotations for partial instances by imposing larger weights on them. Alternatively, various modifications have been introduced to SimpleMIL (Hou et al. 2016; Chen et al. 2019; Chikontwe et al. 2020), aiming at using discriminative instances for training. As shown in Figure 2, these approaches are essentially in the EM framework, training model in M-step and selecting instances in E-step.

Causal Inference in Computer Vision

Causal inference has been introduced to a growing number of computer vision tasks, including class-incremental learning (Hu et al. 2021), long-tailed classification (Tang, Huang, and Zhang 2020), unsupervised representation learning (Wang et al.
Interventional Multi-Instance Learning

Revisit MIL through Causal Inference

Denote \( \{X_i, Y_i\}_{i=1}^t \) as a set of coarsely-labeled bags. The bag \( X_i = \{x_{ij}\}_{j=1}^{N_i} \) contains \( N_i \) unlabeled instances — each instance-level label \( i.e., y_{ij} \in \{0, 1\} \) for \( x_{ij} \), is unavailable. For each \( X_i \), its bag-level label \( Y_i \) is derived under the standard multiple instance assumption (Dietterich, Lathrop, and Lozano-Pérez 1997), \( i.e., Y_i = 1 \) when \( \exists y_{ij} = 1 \) for \( j = 1, \ldots, N_i \), otherwise, \( Y_i = 0 \).

Multi-instance learning aims at training a predictive model based on coarsely-labeled bags. As illustrated in Figure 3(a), we can formulate the MIL framework as a causal graph (aka, a structural causal model or SCM) (Pearl, Glymour, and Jewell 2016), denoted as \( \mathcal{G} = \{N, E\} \). The nodes \( N \) are a set of variables, and the edges \( E \) indicate causal relations in the system, which are shown below:

- \( B \rightarrow X \): We denote \( X \) as the instance and \( B \) as the bag contextual prior. This link reflects the fact that a bag contains multiple instances.
- \( B \rightarrow D \leftarrow X \): We denote \( D \) as the contextual information shared by the instances in the same bag (aka, instance-shared representation), derived based on the bag contextual prior \( B \). This contextual information is naturally encoded by MIL models as manifold bases (Arora et al. 2019), semantic topics (Bau et al. 2017), typical patterns (Zhang, Wu, and Zhu 2018), etc.
- \( X \rightarrow Y \leftarrow D \): We denote \( Y \) as the classification score determined by \( X \) via a direct effect \( X \rightarrow Y \) and an indirect effect \( D \rightarrow Y \). \( X \rightarrow Y \) is obvious, which means the MIL model outputs \( Y \) given \( X \). On the other hand, \( D \rightarrow Y \) indicates that the bag contextual prior affects the instance labels. Note that \( D \rightarrow Y \) always exists in MIL models. Specifically, if \( D \not\rightarrow Y \) in Figure 3(a), the only path that transfers knowledge from \( B \) to \( Y \): \( B \rightarrow X \rightarrow Y \) is blocked by conditioning on \( X \) (d-separation (Pearl, Glymour, and Jewell 2016)), then instance labels are no longer related to bags, which conflicts with the setting of MIL.

Again, take WSI classification as an example: 1) \( B \rightarrow X \): a WSI contains multiple patches, and the patches are of either different tissue types (e.g., mitosis, cellular, neuronal and gland) or diagnoses (e.g., cancer/non-cancer) (Chan et al. 2019). 2) \( B \rightarrow D \leftarrow X \): The patches in a bag share some underlying information or features, e.g., the global low-level features of the bag like colors and textures. Accordingly, \( D \) corresponds to such instance-shared information. 3) \( X \rightarrow Y \leftarrow D \): A MIL model classifies the patches based on both instance-specific and instance-shared representations. Besides WSI classification, other MIL problems like temporal action localization (Narayan et al. 2019) (videos as bags and frames as instances) and weakly-supervised semantic segmentation (Xu et al. 2019) (images as bags and objects as instances) can also be interpreted by the SCM in Figure 3(a).

In our SCM graph, \( B \) confounds \( X \) and \( Y \) via the backdoor path \( X \leftarrow B \rightarrow D \rightarrow Y \) (Pearl 1995), \( i.e., \) predicting all instances in a bag to be the same even if some instances are irrelevant to the prediction. On the other hand, \( X \rightarrow D \rightarrow Y \) is a mediation path (Pearl 2013), which is the key mechanism of MIL. Accordingly, the instance-shared information \( D \) works as a mediator, which encodes the dependencies of instances. Take the objects in an indoor scene as an example. An “indoor” bag tends to contain instances of “TV” rather than “wild animal”, where \( D \) contains the “indoor” semantics, which could narrow the search space for instance prediction \( Y \) by filtering out those that belong to the “outdoor” scene. As detailed in the following sections, the main difference between our IMIL and the existing methods lies in the operations of confounder and mediator.

Proposed Learning Method

An ideal MIL model should capture the true causality between \( X \) and \( Y \). However, from the SCM in Figure 3(a), the conventional correlation of \( P(Y|X) \) fails to do so, because the likelihood of \( Y \) given \( X \) is not only due to \( X \) per se, but also the spurious correlation caused by the confounder \( B \). Therefore, to pursue the true causality between \( X \) and \( Y \), we seek to use the causal intervention \( P(Y|do(X)) \) instead of the likelihood \( P(Y|X) \) for MIL objective. Here, the \( do(\cdot) \) operation is defined as forcibly assigning a specific value to a variable, corresponding to applying random controlled trials (Pearl and Mackenzie 2018). Accordingly, we implement our IMIL framework by the following Expectation-Maximization (EM) algorithm, achieving deconfounded training and discriminative instance selection. In the following content, we denote variable as a capital letter and denote value as lowercase.

M-step: Deconfounded training In M-step, the model is optimized under the physical intervention, which aims to “cut-off” the undesired confounding effect, as shown in Fig-
scores, we approximate the score of each curriculum to set the reference effect adaptively. Our causal intervention adjustment (Pearl, Glymour, and Jewell 2016) to achieve the E whose effects are larger than that of E includes resizing, cropping, and flipping.

In particular, the bag context prior could be instantiated as the boundary separating discriminative instances from non-discriminative ones. Besides, we also consider random rotation, which has been widely used in image recognition tasks (Wan et al. 2013). As demonstrated in the following experiments, such deconfounded training brings significant improvements to all compared methods, which is a practical, generic, and implementation-friendly solution. It should be noted that leveraging the domain knowledge specialized by tasks may help design more effective data augmentation methods, which is left as our future work.

E-step: Discriminative instance selection via total effect

After applying the deconfounded training above in the M-step, we further select discriminative instances in the E-step, suppressing the confounding bias imposed by those non-discriminative instances in the next iteration. In principle, we introduce the Total Effect (TE) defined below as our criterion for instance selection:

\[ TE(Y) = \mathbb{E}[Y|do(X = x)] - \mathbb{E}[Y|do(X = x_0)] = P(1|do(X = x)) - P(1|do(X = x_0)), \]  

(1)

which measures the expected effect on the prediction Y as the instance X changes from x_0 to x. Here, x_0 is the reference instance, whose effect \( \mathbb{E}[Y|do(X = x_0)] \) defines the boundary separating discriminative instances from non-discriminative ones. Accordingly, we select the instances whose effects are larger than that of x_0. Obviously, the estimation of the TE consists of approximating the causal intervention \( P(Y|do(X)) \) and setting the reference effect \( \mathbb{E}[Y|do(X = x_0)] \). In this work, we apply the backdoor adjustment (Pearl, Glymour, and Jewell 2016) to achieve the causal intervention \( P(Y|do(X)) \) and propose a progressive curriculum to set the reference effect adaptively.

At the t-th E-step, given a bag \( \{x_{ij}\}_{j=1}^{N_i} \), we can calculate the score of each \( x_{ij} \) by current model, denoted as \( s^{(t)}(x_{ij}) \). For MIL classification tasks, the score is often derived by a sigmoid or softmax operation. Based on the scores, we approximate \( \mathbb{E}[Y|do(X = x)] \) by an energy-based model (Tang, Huang, and Zhang 2020):

\[ \mathbb{E}[Y|do(X = x)] \propto \frac{S(x_{ij})}{\sum_{i} S(x_{ij})} = E(x_{ij}). \]  

(2)

This model is different from conventional softmax directly derived by the scores. Firstly, the \( S(x_{ij}) \) represents the exponential moving average scores derived by the Temporal Ensembling method (Laine and Aila 2016), which estimates the unnormalized effect of \( do(X = x_{ij}) \). It is calculated as follows:

\[ S(x_{ij}) \leftarrow m S(x_{ij}) + (1 - m) s^{(t)}(x_{ij}). \]  

(3)

This mechanism can be explained as applying momentary interval sampling multiple times, where \( S(x_{ij}) \) is the ensemble of scores and \( m \) is a momentum coefficient. Applying \( S(x_{ij}) \) rather than \( s^{(t)}(x_{ij}) \) helps to enhance the robustness of instance selection (Laine and Aila 2016). Secondly, the denominator of \( E(x_{ij}) \) is the average score of all instances in a bag. It works as a propensity score (Austin 2011), balancing the observational bias of instances, as shown in the scatter plots of Figure 2. Note that our implementation is in the form of inverse probability weighting (Pearl, Glymour, and Jewell 2016) since the confounder \( B \) is unobserved.

For the reference effect, we initialize \( \mathbb{E}[Y|do(X = x_0)] \) as 0 and increase it gradually with the increase of iterations. In particular, we calculate \( E(x_{ij}) \) for each instance and remain the discriminative instances that correspond to the top \( \{ R^{(t)} K \} \) largest values in \( \{ E(x_{ij}) \} \), where \( R^{(t)} = (1 - \tau t) \), \( \tau \) is the decay ratio, and \( K = \sum_{i=1}^{t} N_i \) is the number of instances. Accordingly, the reference effect at the t-th E-step, i.e., \( \mathbb{E}[Y|do(X = x_0^{(t)})] \), is

\[ \max \{ \epsilon : \sum_{i} I(E(x_{ij}) \geq \epsilon) = \lfloor R^{(t)} K \rfloor \}, \]  

(4)

where \( I(\cdot) \) is an indicator function generating 1 if the statement is true. The \( R^{(t)} \) stops updating when the average reweighted score of the \( \tau K \) smallest selected instances exceeds a predefined threshold \( T \), i.e.,

\[ \min_{X \subset X_1} \left\{ \frac{1}{\tau K} \sum_{x \in X} E(x) : |X| = \tau K \right\} \geq T. \]  

(5)

where \( X_1 \) is the set of selected instances. It is noted that the higher the score is, the more likely newly filtered instances are to be discriminative. Such a procedure can effectively prevent MIL models from over-fitting (Wei et al. 2020). Plugging Eq. (2) and Eq. (4) into Eq. (1), we approximate the TE for each instance \( x \) at the t-th E-step as

\[ TE(Y) = E(x) - \mathbb{E}[Y|do(X = x_0^{(t)})]. \]

Justifications

Implementations of Causal Intervention

While causal intervention is agnostic to methods, datasets, and backbones in theory, their implementations are often task-specific in practice. Focusing on the EM framework of MIL, we implement the causal intervention at the M-step and that at the E-step by the physical intervention and the backdoor adjustment method, respectively. The physical intervention increases the diversity of data, which is beneficial for the M-step to avoid the over-fitting problem. However, it removes the confounding bias by introducing new randomness (Sohn et al. 2020) rather than selecting instances. On the contrary, the backdoor adjustment method approximates...
TE to select discriminative instances. As shown in Figure 4, approximating TE helps to remove the undesired confounder bias (i.e., cut off the backdoor path \(X \leftarrow B \rightarrow D \rightarrow Y\)) and keep the mediator effect (i.e., reserve the mediator path \(X \rightarrow D \rightarrow Y\)), which is more suitable for the E-step.

### Multi-class Multi-label MIL Problems

Our IMIL can be easily applied to multi-class multi-label MIL problems, as long as we take each class as a classical (binary) MIL problem. In such situations, the concept of “discriminative” varies from class, and the non-discriminative instance in one class may provide reliable supervision as a negative instance for another class. Therefore, we may select instances in a relatively conservative manner. Specifically, if one instance is not selected for one class, we only reduce the significance of its loss in this class while maintaining its significance for the other classes. This procedure can be understood as a more fine-grained soft version of our selection method.

### Connections with Existing MIL Methods

It should be noted that our IMIL framework provides a new way to analyze existing state-of-the-art MIL methods. Specifically, Table 1 makes a comparison for various MIL methods from the viewpoint of causal intervention. Essentially, most existing MIL methods can be categorized into three classes according to their instance selection strategies. The method in the first class is the SimpleMIL (Cheplygina et al. 2017), which simply uses all instances without causal intervention; The methods in the second class select instances by calculating the Natural Direct Effect (NDE) shown in Figure 4. In particular, \(\text{NDE}(Y) = \mathbb{E}[Y_{d_0}|do(X = x)] - \mathbb{E}[Y_{d_0}|do(X = x_0)]\), where \(Y_{d_0}\) is the counterfactual output achieved under the condition of \(do(X = x_0)\). Because the NDE completely removes the entire effect of \(D\), which may lose some information beneficial for the learning problem, the MIL methods in this class often require some external information as compensations. The representative methods include Top-kMIL (Chikontwe et al. 2020), RCEMIL (Chen et al. 2019) and PatchCNN (Hou et al. 2016). These methods select instances by comparing the scores of instances from the same bag, which is actually an intervention on \(D\) that forces the mediator-specific effect to be the same. The \(x_0\)

**Table 1: The differences among various MIL methods.**

<table>
<thead>
<tr>
<th>MIL Simple</th>
<th>Patch</th>
<th>RCE</th>
<th>Top-k</th>
<th>Semi</th>
<th>IMIL</th>
</tr>
</thead>
<tbody>
<tr>
<td>(do(D))</td>
<td>-</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>(do(X))</td>
<td>-</td>
<td>-</td>
<td>✓</td>
<td>✓</td>
<td>-</td>
</tr>
<tr>
<td>Effect</td>
<td>NDE</td>
<td>NDE</td>
<td>NDE</td>
<td>TE</td>
<td>TE</td>
</tr>
</tbody>
</table>

**Table 2: Summary of the WSI datasets.** N/A: not available.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Class</th>
<th>Instance Size</th>
<th># of bags</th>
<th># of instances</th>
<th># of instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>DigestPath</td>
<td>Malignant Normal</td>
<td>512</td>
<td>250</td>
<td>10,133</td>
<td>N/A</td>
</tr>
<tr>
<td>Camelyon16</td>
<td>Metastases Normal</td>
<td>256</td>
<td>40</td>
<td>64,430</td>
<td>60,545</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>410</td>
<td>33,110</td>
<td>60,545</td>
</tr>
</tbody>
</table>

**Experiments**

**Pathological Image Analysis**

To demonstrate the effectiveness of our IMIL method, we apply it to the WSI classification problem and compare it with state-of-the-art MIL methods. In particular, we consider the classification of colonoscopy tissues (malignant v.s. normal) and the classification of lymph node sections (metastases v.s. normal). The datasets used in our experiments are DigsetPath (Li et al. 2019a)² and Camelyon16 (Bejnordi et al. 2017)³, both of which have bag-level and instance-level labels for each image and its patches, respectively. Specifically, we apply Otsu’s method (Otsu 1979) to remove the background of the images and extract non-overlapped patches from the foreground regions. The statistics of these two datasets are summarized in Table 2.

The competitors of our method include an Oracle model with full supervision at the instance-level and the state-of-the-art MIL methods like the SimpleMIL in (Cheplygina et al. 2017), the PatchCNN in (Hou et al. 2016), the RCEMIL trained by RCE loss (Chen et al. 2019), the SemiMIL trained with both bag-level and partial instance-level labels, and the Top-kMIL trained by Top-k selection (Chikontwe et al. 2020). For PatchCNN, SemiMIL and RCEMIL, some external information should be provided: the specific threshold for each bag, partial instance-level labels, and two statistical values for the re-weighting scheme. For other methods, including ours, only the bag-level labels are used for training.

³https://camelyon16.grand-challenge.org/
For a fair comparison, all the methods use ResNet-18 (He et al. 2016) as their backbone models. Adam optimizer is used with an initial learning rate of 0.001, and the batch size is set to 64. We run 50 epochs in total and decay the learning rate with the cosine decay schedule (Loshchilov and Hutter 2016). For our method, the hyperparameters are $m = 0.5$, $\tau = 0.05$ and $T = 0.95$ by default. We evaluate the instance-level performance of each method based on 5-fold cross-validation, and the measurements include Area Under Curve (AUC), accuracy (ACC), F1-score, recall (REC) and precision (PRE). The numerical comparisons for various methods on the two datasets are shown in Table 3. We can find that the causal intervention consistently improves the methods in all settings, indicating that causal intervention is agnostic to methods and datasets.

For the methods using external information, the finer granularity the supervision used, the better performance is achieved. However, their performance is also highly dependent on the quality of external labels, which decreases dramatically given noisy external labels (denoted by *). Note that the results of the oracle model should indicate the upper bound of obtainable performance, while its low recall on Camelyon16 is mainly due to the highly imbalanced ratio of the data. Among the methods without external information, our IMIL obtains superior performance. It consistently improves SimpleMIL on most measurements by non-trivial margins. Especially, IMIL remarkably promotes precision by 15.3% and 16.33% on DigestPath and Camelyon16, respectively, indicating a substantial reduction over false positives. Although the method of Top-$k$ MIL performs well on DigestPath, it receives extremely low recall (15.28% without causal intervention) on Camelyon16 because the setting of $k$ is sensitive to the change of bag size. Additionally, the top-$k$ selection may degenerate into max-max selection criteria, which tends to have relatively low recall and high specificity (Xu et al. 2019). On the contrary, our method can adaptively select discriminative instances instead of setting a fixed number.

### Table 3: Numerical results (%) on dataset of DigestPath and Camelyon16. ‘+CI’ means that causal intervention is applied in the M-step. ‘*’ means the external information is噪声.

<table>
<thead>
<tr>
<th>Method</th>
<th>DigestPath</th>
<th>Camelyon</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AUC</td>
<td>ACC</td>
</tr>
<tr>
<td>Oracle</td>
<td>93.26</td>
<td>88.51</td>
</tr>
<tr>
<td>simpleMIL</td>
<td>89.52+3.16</td>
<td>90.66±2.15</td>
</tr>
<tr>
<td>R-CEMIL</td>
<td>87.20</td>
<td>82.75</td>
</tr>
<tr>
<td>PatchCNN</td>
<td>91.09</td>
<td>83.46</td>
</tr>
<tr>
<td>SimpleMIL</td>
<td>94.53+3.44</td>
<td>88.28±4.82</td>
</tr>
<tr>
<td>Top-$k$ MIL</td>
<td>91.84-2.69</td>
<td>82.13±6.15</td>
</tr>
<tr>
<td>Oracle</td>
<td>91.94</td>
<td>87.68</td>
</tr>
<tr>
<td>SimpleMIL</td>
<td>92.81-1.59</td>
<td>85.88±4.22</td>
</tr>
<tr>
<td>R-CEMIL</td>
<td>91.05+2.59</td>
<td>79.70±2.50</td>
</tr>
<tr>
<td>PatchCNN</td>
<td>93.01-4.44</td>
<td>85.75±4.78</td>
</tr>
<tr>
<td>IMIL (Ours)</td>
<td>88.46</td>
<td>77.20</td>
</tr>
<tr>
<td>Top-$k$ MIL</td>
<td>89.57</td>
<td>80.97</td>
</tr>
<tr>
<td>SimpleMIL</td>
<td>93.01-4.44</td>
<td>85.75±4.78</td>
</tr>
<tr>
<td>IMIL (Ours)</td>
<td>88.16</td>
<td>81.39</td>
</tr>
</tbody>
</table>

Figure 5: AUC under different composition of data augmentations, where −/+ means one individual data augmentation is removed/added.

### Further Analysis

#### Data Augmentations
To quantitatively assess the contributions of different data augmentations, we remove/add individual data augmentation in the M-step on three methods (i.e., Oracle, SimpleMIL, and our IMIL). For these methods, removing/adding a significant augmentation method is expected to harm/improve their performance. Figure 5 indicates that "color jittering" is the most important for IMIL and SimpleMIL — the instances in the same bag share both labels and staining conditions, thus jittering could prevent their models from exploiting this spurious correlation. For the oracle model, the "color jittering" is unimportant since the instances are fully-supervised, and the staining condition will cause less confounding bias. Instead, the models may over-fit the co-occurrence of tissues, thus "random resizing and cropping" is necessary for removing these biases. For other augmentations, "grayscale conversion" and "Gaussian blurring" achieve the causal intervention at the cost of hurting the characteristics (the HE staining and the resolution) of WSIs, while "flipping" and "rotation" bring limited improvements because WSIs have no dominant orientation.
We present qualitative results of instance selection, while a large one may be too aggressive. Overall, our IMIL is robust to the hyperparameters.

Robustness to Hyperparameters For those important hyperparameters of our method, we change one of them in a range and fix the others to their default values. The performance of our IMIL under different configurations is shown in Table 4. For the momentum used in Temporal Ensembling, the best performance is achieved when $m = 0.75$. This result reflects that a relatively large momentum is beneficial for robust instance selection. For the threshold used in selecting curriculum, we set $T$ around 1 for the assumption that scores of positive and negative instances can be separated by average scores in most case, as shown in Figure 2. Intuitively, higher threshold will keep less instances selected and result in lower recall ($T = 0.05$). For the step used in selecting curriculum, a small $\tau$ results in a slow process of instance selection, while a large one may be too aggressive.

Table 4: The robustness of our IMIL to its hyperparameters.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Settings</th>
<th>AUC</th>
<th>Acc</th>
<th>F1</th>
<th>REC</th>
<th>PRE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m = 0$</td>
<td>92.18</td>
<td>86.69</td>
<td>73.55</td>
<td>79.25</td>
<td>68.80</td>
<td></td>
</tr>
<tr>
<td>$m = 0.25$</td>
<td>91.97</td>
<td>86.72</td>
<td>73.37</td>
<td>78.16</td>
<td>69.25</td>
<td></td>
</tr>
<tr>
<td>Momentum $m = 0.5$</td>
<td>92.06</td>
<td>86.84</td>
<td>73.74</td>
<td>78.77</td>
<td>69.60</td>
<td></td>
</tr>
<tr>
<td>$m = 0.75$</td>
<td>92.47</td>
<td>87.32</td>
<td>74.68</td>
<td>79.32</td>
<td>70.70</td>
<td></td>
</tr>
<tr>
<td>$m = 0.9$</td>
<td>91.60</td>
<td>85.56</td>
<td>72.88</td>
<td>79.75</td>
<td>68.07</td>
<td></td>
</tr>
<tr>
<td>Threshold $T = 0.9$</td>
<td>91.67</td>
<td>86.69</td>
<td>72.66</td>
<td>76.34</td>
<td>69.55</td>
<td></td>
</tr>
<tr>
<td>$T = 0.95$</td>
<td>92.06</td>
<td>86.84</td>
<td>73.74</td>
<td>78.77</td>
<td>69.60</td>
<td></td>
</tr>
<tr>
<td>$T = 1$</td>
<td>92.33</td>
<td>87.35</td>
<td>74.34</td>
<td>77.94</td>
<td>71.46</td>
<td></td>
</tr>
<tr>
<td>$T = 1.05$</td>
<td>91.28</td>
<td>86.79</td>
<td>72.23</td>
<td>74.12</td>
<td>70.49</td>
<td></td>
</tr>
<tr>
<td>$\tau = 0.025$</td>
<td>91.99</td>
<td>86.90</td>
<td>73.88</td>
<td>78.02</td>
<td>70.46</td>
<td></td>
</tr>
<tr>
<td>Step $\tau = 0.05$</td>
<td>92.06</td>
<td>86.84</td>
<td>73.74</td>
<td>78.77</td>
<td>69.60</td>
<td></td>
</tr>
<tr>
<td>$\tau = 0.075$</td>
<td>92.46</td>
<td>87.43</td>
<td>74.56</td>
<td>78.20</td>
<td>71.46</td>
<td></td>
</tr>
</tbody>
</table>

Qualitative Results We present qualitative results of IMIL in two aspects: the selecting curriculum procedure and the patch classification on new WSIs. In Figure 6(a), discriminative instances are gradually selected, validating our initial assertions that the average scores of bags naturally can serve as propensity scores. Though the achieved supervision is not perfectly clean, it is worth noting that we do not need any external supervision. In Figure 6(b), the heatmaps indicate the regions of high tumor probability. Notably, our IMIL can accurately distinguish tumors from normal tissues and considerably reduce false positives compared to SimpleMIL, suggesting the effectiveness of the proposed method.

Bag-Level Classification on Pascal VOC To demonstrate the universality of our IMIL method, besides WSI classification problems, we further test it on the PASCAL VOC 07 dataset (Everingham et al. 2015). This dataset contains 9963 natural images of 20 categories, which is challenging as the appearances of objects are diverse. Following (Wu et al. 2015b), we take each image as a bag and adopt region proposal methods, e.g., Region Proposal Network (RPN) (Ren et al. 2015), to generate instances. Since the instance-level labels are unavailable in VOC, we only deploy the methods free of extra information and equip them with three MIL pooling operators, i.e., max pooling, mean pooling and voting, for bag classification. For all the methods, we report its mean average precision (mAP) on the test set in Table 5. The proposed causal intervention steadily improves all compared methods, where our IMIL achieves the best performance on bag-level prediction, demonstrating the proposed framework’s stability and effectiveness.

Table 5: The mean average precision over Pascal VOC 07.

<table>
<thead>
<tr>
<th>MIL</th>
<th>Aggregator</th>
<th>max</th>
<th>mean</th>
<th>voting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple</td>
<td>+CI</td>
<td>63.89</td>
<td>68.2</td>
<td>65.46</td>
</tr>
<tr>
<td>Top-k</td>
<td>+CI</td>
<td>74.8+10.91</td>
<td>77.97+9.77</td>
<td>75.46+10.0</td>
</tr>
<tr>
<td>IMIL</td>
<td>+CI</td>
<td>75.78+10.78</td>
<td>78.34+9.69</td>
<td>75.53+9.87</td>
</tr>
</tbody>
</table>

Conclusion

We present a novel interventional multi-instance learning (IMIL) method to suppress the negative influence of bag contextual prior to instance-level prediction. In particular, we propose a causal graph for MIL and equip the EM-based MIL paradigm with causal intervention, combining the training process with data augmentations and adaptive instance selection. Experimental results show that our IMIL achieves promising performance on various computer vision tasks. In the future, we plan to design more task-specific data augmentation methods to improve the physical intervention strategy. Moreover, a systematic comparison will be considered for various causal intervention methods in MIL tasks.

Acknowledgments

Dr. Yi Xu was supported in part by NSFC 62171282, Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), 111 project BP0719010, and SJTU Science and Technology Innovation Special Fund ZHK2018ZDA17. Dr. Hongteng Xu was supported in part by NSFC 62106271 and Intelligent Social Governance Platform, Major Innovation & Planning Interdisciplinary Platform for the “Double-First Class” Initiative, RUC. He also thanks the support of Public Policy and Decision-making Research Lab of RUC and JD Explore Academy. Authors thank Zengchao Xu for inspiring discussions.
References


