COMPACT AND DE-BIASED NEGATIVE INSTANCE EMBEDDING FOR MULTI-INSTANCE LEARNING ON WHOLE-SLIDE IMAGE CLASSIFICATION

Joohyung Lee⋆, Heejeong Nam⋆, Kwanhyung Lee⋆, Sangchul Hahn⋆

⋆AITRICS, Republic of Korea

ABSTRACT

Whole-slide image (WSI) classification is a challenging task because 1) patches from WSI lack annotation, and 2) WSI possesses unnecessary variability, e.g., stain protocol. Recently, Multiple-Instance Learning (MIL) has made significant progress, allowing for classification based on slide-level, rather than patch-level, annotations. However, existing MIL methods ignore that all patches from normal slides are normal. Using this free annotation, we introduce a semi-supervision signal to de-bias the inter-slide variability and to capture the common factors of variation within normal patches. Because our method is orthogonal to the MIL algorithm, we evaluate our method on top of the recently proposed MIL algorithms and also compare the performance with other semi-supervised approaches. We evaluate our method on two public WSI datasets including Camelyon-16 and TCGA lung cancer and demonstrate that our approach significantly improves the predictive performance of existing MIL algorithms and outperforms other semi-supervised algorithms. We release our code at https://github.com/AITRICS/pathology_mil.

Index Terms— whole-Slide Image, WSI classification, Multiple-Instance Learning, Semi-Supervised Learning

1. INTRODUCTION

Whole slide images (WSI) are digitized histology slides that preserve the bountiful information of original histology [1, 2]. With WSI, pathologists no longer use glass slides with a microscope but can apply various computing tools to process the digitized image [3]. WSI has not only improved the workflow and diagnostic procedure of pathologists but also enabled disease analysis, e.g., cancer prognosis analysis, to benefit from deep learning technology. However, two main difficulties hinder the application of the conventional deep learning methods to whole slide image analysis.

First, WSIs are extremely large in size, often reaching up to 10^{10} pixels [1]. Due to its exceptionally large size, existing methods usually employ patch-based processing - a WSI is split into small image patches which are then processed using the image encoder. However, a thorough examination of the whole slide image for a patch/pixel-level annotation is nearly infeasible due to the exceptionally large size of WSI,

and therefore, a label for each image patch is usually lacking. Recently, multiple-instance learning (MIL) has been drawing attention in disease analysis of whole slide images since it exempts costly pixel-level annotation, which was required for traditional machine learning (ML) approaches for WSI analysis [4]. Among various approaches within the MIL algorithm, attention-based MIL has shown promising results recently, which aggregates all instance embeddings into a single bag-level embedding using an attention module [4, 5, 6].

Second, WSIs possess wide variations, such as differences in tumor types, pen ink, and staining protocols [7, 8, 9, 10]. Though some variations can be informative, some variations are not and may even create a spurious correlation to the label. Lin *et al.* [9] has introduced a concept of 'bag prior' and defined it as 'an instance(patch)-shared information per bag (WSI) but irrelevant to the label', which therefore can cause the spurious correlation between instances and labels. In this study, we suggest improving the predictive performance of existing MIL algorithms by reducing the inter-slide bias by implementing a synchronized instance embedding center. The method and the result are depicted in Fig. 1 and 2-(a), respectively.

The lack of local annotation for WSI is no longer problematic due to the widespread usage of the MIL algorithm. However, it has to be noted that WSI implicitly provides local annotation for negative slides; every tissue region from negative slides is negative. Therefore, even without explicit local annotation, we have access to normal-label (disease-negative) patches, which are disregarded in MIL. We aim to use this implicit information from WSI.

Such circumstance has been studied in one-class classification (OCC), where only a single class data is present during the training phase; data from another class, i.e., anomalous data, can appear during the test phase. Various approach has been introduced including the generative approach, synthesizing anomalies, and embedding-based approach [11]. Motivated by the embedding-based approach, we hypothesize that mapping all normal patches into a compact space can not only reduce the aforementioned inter-slide bias but also facilitate the model to capture the common factors of variation within the normal patches. Capturing common factors within normal patches can help distinguish normal, i.e. negative, and positive patches. We thus aim to demonstrate that creating com-
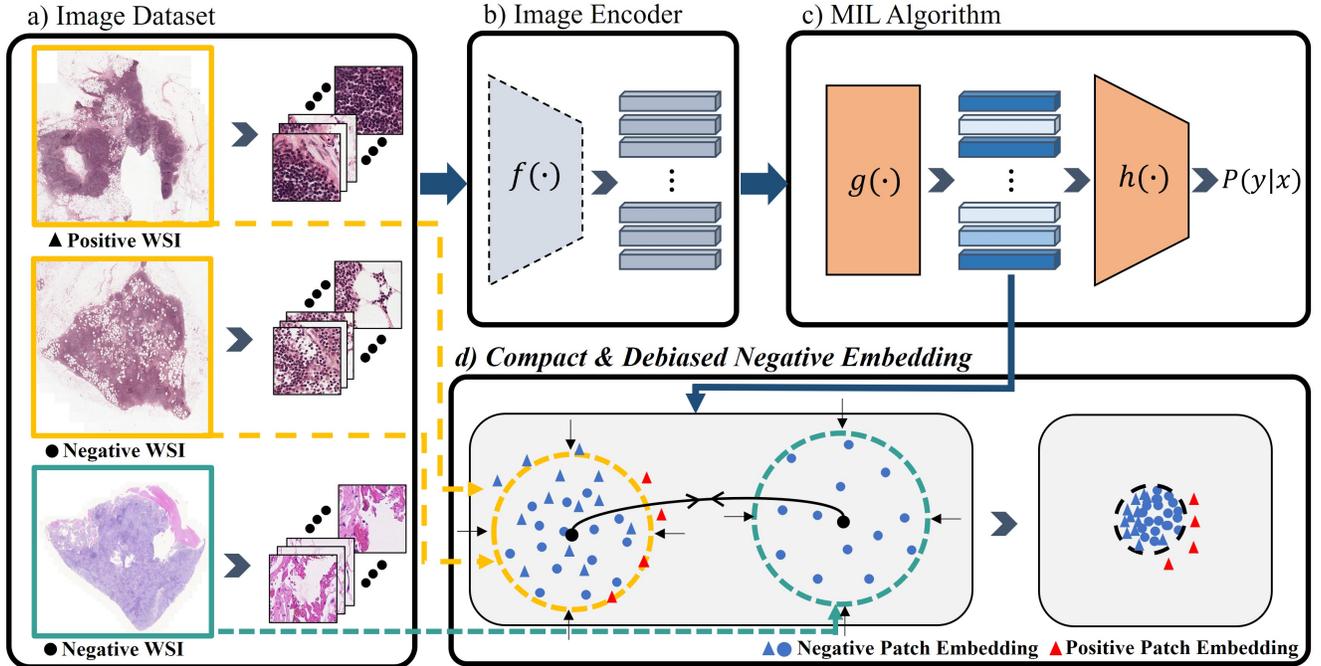
Fig. 1. Overview of our proposed **Compact and Debiased Negative Embedding** attached to the MIL algorithm. In this figure, the staining protocol is exemplified as the 'bag-prior'; WSIs with darker stains are in yellow boxes whereas the green box indicates the WSI with lighter staining. Our method aims 1) to reduce the 'bag-prior' by synchronizing the center of all patch embeddings from negative WSIs and 2) to impose a compact embedding for all patch embeddings from negative WSIs, as indicated by the black arrows. Note that triangles and dots in (d) are instance embeddings from different WSI indicated in (a).

pact normal patch embedding improves the predictive performance of existing MIL algorithms.

In this study, we aim to enhance the slide-level predictive accuracy of existing MIL methods using patches from the disease-negative slides. Specifically, we propose to minimize the standard deviation of all instance embeddings (before the attention aggregator) from negative WSI using a single learnable negative instance mean embedding vector. We hypothesize that compact negative representation will lead to more precise attention thus leading to the improved predictive performance of MIL algorithm.

We evaluate our method on two public WSI datasets, i.e., Camelyon-16 and TCGA lung cancer, to demonstrate that our approach improves the slide-level prediction of the existing MIL algorithms. We also compare the performance gain of our algorithm with other semi-supervised algorithms in WSI classification using MIL as well.

2. METHOD

In this paper, we propose a semi-supervision signal that utilizes all patches from disease-negative slides to improve the predictive performance of the MIL model. Our aim is 1) to reduce the bias of each WSI, e.g., caused by varying staining protocol, and 2) to facilitate the mapping module in MIL al-

gorithms, i.e., $g(\cdot)$ in Fig. 1, to capture the common factors of variation within the normal patches, which can help assign a more precise attention score to each instance in attention-based MIL. Note that our method is orthogonal to the existing MIL algorithms and thus can be attached to any existing MIL algorithms to improve the predictive performance.

Definition 1 (WSI Dataset). *We consider a WSI dataset $\mathcal{S}$ to consist of N whole slide images. Elements of $\mathcal{S}$ are tuples, i.e. $\mathcal{S} := \{(s_1, y_1^{wsi}), ..., (s_N, y_N^{wsi})\}$, where s_i and y_i^{wsi} denotes the ith whole slide image and its disease class label.*

Definition 2 (Patch Dataset). *i^{th} whole slide image s_i consists of K tuples after preprocessing, i.e., $s_i := \{(p_{ik}, y_{ik}^{patch})\}_{k=1}^K$. p_{ik} and y_{ik}^{patch} denote the extracted image patch and its label which is only known when the corresponding WSI is disease-negative, i.e., $y_{ik}^{patch} = 0 \ \forall k$ when $y_i^{wsi} = 0$. The number of image patches K varies per slide s_i. We denote the encoded image patch $x_{ik} := f(p_{ik})$ and $x_i := \{x_{ik}\}_{k=1}^K$ where $f(\cdot)$ is the offline image encoder in Fig. 1.*

Revisiting Multiple Instance Learning for Whole Slide Image Classification. In general, as illustrated in Fig.1, the WSI classification pipeline usually consists of two stages: 1) image encoder $f(\cdot)$, and 2) MIL module (Fig. 1-c). [12]. For the image encoder, most recent studies utilize offline

Table 1. Data statistics of Camelyon-16 and TCGA-Lung

Dataset		Training	Test	Total
Camelyon-16	# WSI	270	129	399
	# Patch	2,428,707	1,174,691	3,603,398
TCGA-Lung	# WSI	751	248	999
	# Patch	12,048,148	3,156,655	15,204,803

Table 2. Performance comparison between baseline MIL models and MIL models with our proposed method.

Dataset		Camelyon-16		TCGA-Lung	
Method		AUC	ACC	AUC	ACC
DSMIL[4]		85.08	86.82	97.47	93.18
	+Ours	**88.11**	**87.75**	**97.56**	**93.84**
	Δ	+3.03	+0.93	+0.09	+0.66
DTFD-MIL[5]		86.28	85.74	97.79	93.93
	+Ours	**90.16**	**88.37**	**97.88**	**94.41**
(*AFS*)	Δ	+3.88	+2.63	+0.09	+0.48
ABMIL[6]		82.61	86.2	96.4	89.19
	+Ours	**89.24**	**87.6**	**96.58**	**97.7**
(*Attention*)	Δ	+6.63	+1.4	+0.18	+8.51
ABMIL[6]		85.71	86.36	96.53	**97.38**
	+Ours	**88.77**	**87.6**	**96.99**	90.24
(*GatedAttention*)	Δ	+3.06	+1.24	+0.46	-7.14

pre-trained image encoder, e.g., using ImageNet, which can reduce the computational cost of model training and memory consumption [12, 7]. We use a self-supervised pre-trained ResNet-50 using ImageNet [13] and DSMIL pre-trained ResNet-18 [4] for Camelyon-16 and TCGA-Lung cancer classification tasks, respectively.

MIL module (Fig. 1-c) usually consists of two stages: 1) mapping function $g(\cdot)$ and instance embedding aggregator $h(\cdot)$, which, in recent popular attention-based MIL, aggregate each instance embedding using their respective attention score using querying function [4, 5, 6]. In this study, we aim to constrain $g(\cdot)$ to yield compact embedding for all disease-negative patches.

Compact Negative Embeddings with Debiased Bag Prior. We hypothesize that both 1) learning compact negative instance representations from disease-negative bags and 2) reducing the 'bag prior' [14, 15] can improve the predictive performance of the MIL algorithm. Note that Fig. 1 depicts the exemplary 'bag prior' caused by the staining protocol; homogeneous/heterogeneous staining protocol can lead to more similar/less similar image patch embedding, and it can be wrongly exploited by the MIL model and thus may create a spurious correlation between unnecessary image context and the label.

Considering that all patches from the negative slide are disease-negative and from the positive slide are mostly negative as well [4], we were motivated by the approach from the anomaly detection community to impose a compact description of normal data [14, 15]. To this end, we design a supple-

mentary loss function to minimize the standard deviation of patch embeddings $g(x_i)$ when $y_i^{wsi} = 0$ (negative WSIs) as well as to synchronize the center of patch embeddings from every negative slide. Specifically, we used a single linear projection layer $l(\cdot)$ to transform $g(x_i)$ into a M-dimension representation and impose compact embeddings for all negative slides (we used $M = 128$ throughout the study). We also jointly train M-dimension distribution center of the disease-negative instance embedding $\hat{\mu}$ to reduce the 'bag prior' by minimizing the standard deviation of instance embeddings from all negative WSI using a single distribution center. The equation can be shown in Eq. 1:

$$\mathcal{L}_{neg} = \frac{1}{M} \sum_m^M \sqrt{\frac{1}{K-1} \sum_{k=1}^K (g(x_ik) - \hat{\mu})^2} \quad (1)$$

Note that $\hat{\mu}$ is a learnable parameter shared by all WSI slides to synchronize the center of negative embedding. Our algorithm is not bound to a binary classification task but is applicable to multi-class classification tasks as well. In a multi-class classification scenario, we learn multiple embedding centers with multiple linear projection layers $l(\cdot)$ (equal to the class number). To avoid degenerate solutions, e.g., zeros for all parameters in $l(\cdot)$, we increase the standard deviation when the standard deviation is smaller than the threshold thr. The equation can be shown in Eq. 2:

$$\mathcal{L}_{pos} = \frac{1}{M} \sum_m^M ReLU\left(thr - \sqrt{\frac{1}{K-1} \sum_{k=1}^K (g(x_{ik}) - \hat{\mu})^2}\right) \quad (2)$$

As a result, the overall loss can be described as below. Throughout this paper, we call our method CDNE, i.e. compact and debiased negative embedding:

$$\mathcal{L}_{overall} = \mathcal{L}_{MIL} + \lambda_{neg}\mathcal{L}_{neg} + \lambda_{pos}\mathcal{L}_{pos} \quad (3)$$

3. EXPERIMENT

Dataset. In this study, we employ two datasets for histopathological analysis: Camelyon-16 and TCGA-Lung (Table 1). The Camelyon-16 dataset comprises 399 WSIs, with 270 allocated for training and 129 for testing, totaling over 3.6 million patches. On the other hand, the TCGA-Lung dataset features 999 WSIs, with 751 designated for training and 248 for testing, culminating in more than 15.2 million patches. As a preprocessing, we segmented both datasets into smaller patches with a fixed mpp of 0.5 and disregarded background patches.
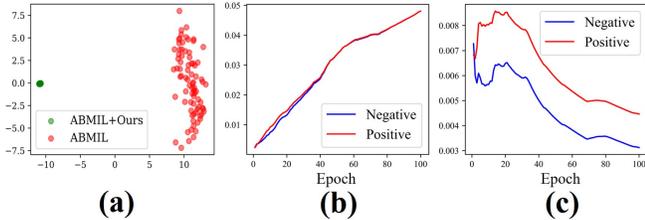
Fig. 2. (a) Instance embedding center per normal test WSIs (80 dots in both green and red); Mean standard deviation of instance embeddings per validation WSIs changes in ABMIL (Gated Attention) without ours (b); and with ours (c).
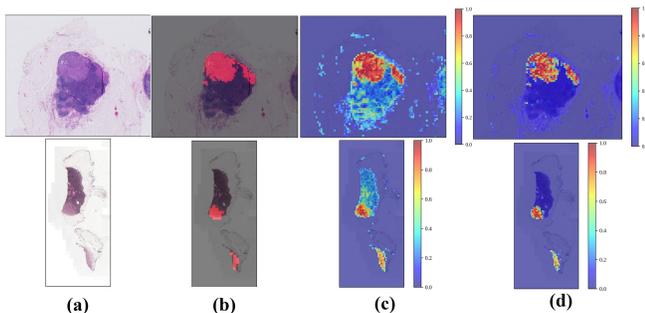

Fig. 3. Attention map comparison with/without our method on ABMIL (Gated Attention) with 2 test WSIs. (a) WSI without annotation; (b) WSI with annotation; (c) heatmap from ABMIL (Gated Attention) baseline; (d) heatmap from Ours + ABMIL (Gated Attention) baseline

For all performance evaluations, we use the averaged area under the receiver operating characteristic (AUROC) and prediction accuracy from 5-fold cross-validation (CV) using a test set.

CDNE increases the performance of MIL. We implemented four recent MIL algorithms and examined the efficacy of our approach on top of them: DSMIL[4], DTFD-MIL[5], ABMIL-Attention, and ABMIL-Gated Attention [6]. To assess the performance of our method, we benchmark the performance of these MIL models both with and without employing our CDNE. As illustrated in Table 2, the application of CDNE significantly boosts both the AUROC and accuracy metrics for the Camelyon-16 dataset. In the case of the TCGA-Lung dataset, where baseline MIL models already demonstrate strong performance, CDNE still manages incremental improvements in predictive accuracy, with the exception of the ABMIL-Gated Attention in accuracy metric. Note that both pulling through negative embeddings (eq.1) and pushing through positive embeddings (eq.2) contribute to the performance increase, and there exists an optimal loss weight for both losses as described in Table. 4.

CDNE outperforms other semi-supervised method. Our approach can be categorized as semi-supervised learning

Table 3. Performance comparison between baseline DSMIL[4] and DSMIL with various SSL methods.

Method	Dataset			
	Camelyon-16		TCGA-Lung	
	AUC	ACC	AUC	ACC
DSMIL[4]	85.08	86.82	97.47	93.17
+Ours	**88.11**	**87.75**	**97.56**	**93.84**
Δ	+3.03	+0.93	+0.09	+0.67
+DivDis[16]	82.29	85.11	97.5	92.98
Δ	-2.79	-1.71	+0.03	-0.19
+Psuedo-Label[17]	77.36	81.6	97.16	92.1
Δ	-7.72	-5.22	-0.31	-1.07

Table 4. Ablation study on $\lambda_{Negative}$ and $\lambda_{Positive}$ with DSMIL+ours in Camelyon-16 test set.

$\lambda_{Negative}$	0	1	10	100	1000
AUC	85.87	86.80	88.11	70.71	64.95
ACC	87.44	86.98	87.75	69.77	61.71
$\lambda_{Positive}$	0	0.3	3	30	300
AUC	86.93	81.78	88.11	85.0	84.45
ACC	87.13	83.26	87.75	86.67	86.51

(SSL) in that it utilizes the implicit patch labels of disease-negative WSI; all patches from the negative slides are negative. Consequently, we compare the performance gain with other semi-supervised approaches. Specifically, we compare the predictive performance of MIL when attached with CDNE (ours), Myronenko [17], and DivDis [16] in Dsmil setting [4]. As described in Table 3, our approach outperforms other alternatives that can utilize the patch label (negative) from negative WSI.

CDNE de-biases negative embeddings and learn better discriminative feature to distinguish positive and negative WSI. As shown in Fig. 2-(a), our method effectively reduces the inter-slide bias when attached to the existing MIL algorithm. Moreover, our compact negative embedding approach creates distinct standard deviations between positive and negative WSIs as described in figure 2-(b,c). Our compact negative embedding method also leads to a more accurate attention map as described in Fig. 3. Note that our method significantly reduces the false positives in Fig. 3.

4. CONCLUSION

We hypothesize and demonstrate that reducing the inter-slide bias and learning a compact negative embedding improve the predictive performance of the MIL algorithm. To this end, we use the implicit labels of negative WSIs and compress their instance embeddings to a single negative center. Through experiments, we confirm that our method effectively reduces the inter-slide bias and creates compact negative embeddings by showing the standard deviation difference between positive and negative WSIs (Fig. 2).

5. REFERENCES

[1] Yuqi Chen, Juan Liu, Zhiqun Zuo, Peng Jiang, Yu Jin, and Guangsheng Wu, "Classifying pathological images based on multi-instance learning and end-to-end attention pooling," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[2] 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," *Advances in neural information processing systems*, vol. 34, pp. 2136–2147, 2021.

[3] Gaojie Li, Qing Liu, Haotian Liu, and Yixiong Liang, "A novel transformer-based pipeline for lung cytopathological whole slide image classification," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[4] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 14318–14328.

[5] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18802–18812.

[6] Maximilian Ilse, Jakub Tomczak, and Max Welling, "Attention-based deep multiple instance learning," in *International conference on machine learning*. PMLR, 2018, pp. 2127–2136.

[7] Xiyue Wang, Jinxi Xiang, Jun Zhang, Sen Yang, Zhongyi Yang, Ming-Hui Wang, Jing Zhang, Wei Yang, Junzhou Huang, and Xiao Han, "Scl-wc: Cross-slide contrastive learning for weakly-supervised whole-slide image classification," *Advances in neural information processing systems*, vol. 35, pp. 18009–18021, 2022.

[8] Tiancheng Lin, Zhimiao Yu, Hongyu Hu, Yi Xu, and Chang-Wen Chen, "Interventional bag multi-instance learning on whole-slide pathological images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19830–19839.

[9] Tiancheng Lin, Hongteng Xu, Canqian Yang, and Yi Xu, "Interventional multi-instance learning with deconfounded instance-level prediction," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, vol. 36, pp. 1601–1609.

[10] Julianna D Ianni, Rajath E Soans, Sivaramakrishnan Sankarapandian, Ramachandra Vikas Chamarthi, Devi Ayyagari, Thomas G Olsen, Michael J Bonham, Coleman C Stavish, Kiran Motaparthi, Clay J Cockerell, et al., "Tailored for real-world: a whole slide image classification system validated on uncurated multi-site data emulating the prospective pathology workload," *Scientific reports*, vol. 10, no. 1, pp. 3217, 2020.

[11] Zhikang Liu, Yiming Zhou, Yuansheng Xu, and Zilei Wang, "Simplenet: A simple network for image anomaly detection and localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 20402–20411.

[12] Chao Tu, Yu Zhang, and Zhenyuan Ning, "Dual-curriculum contrastive multi-instance learning for cancer prognosis analysis with whole slide images," *Advances in Neural Information Processing Systems*, vol. 35, pp. 29484–29497, 2022.

[13] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," *Advances in neural information processing systems*, vol. 33, pp. 9912–9924, 2020.

[14] Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft, "Deep one-class classification," in *International conference on machine learning*. PMLR, 2018, pp. 4393–4402.

[15] Lukas Ruff, Robert A Vandermeulen, Nico Görnitz, Alexander Binder, Emmanuel Müller, Klaus-Robert Müller, and Marius Kloft, "Deep semi-supervised anomaly detection," *arXiv preprint arXiv:1906.02694*, 2019.

[16] Yoonho Lee, Huaxiu Yao, and Chelsea Finn, "Diversify and disambiguate: Learning from underspecified data," *arXiv preprint arXiv:2202.03418*, 2022.

[17] 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 *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 329–338.