

---

# Retrieval-Augmented Multiple Instance Learning

---

Anonymous Author(s)

Affiliation

Address

email

## Abstract

Multiple Instance Learning (MIL) is a crucial weakly supervised learning method applied across various domains, e.g., medical diagnosis based on whole slide images (WSIs). Recent advancements in MIL algorithms have yielded exceptional performance when the training and test data originate from the same domain, such as WSIs obtained from the same hospital. However, this paper reveals a performance deterioration of MIL models when tested on an out-of-domain test set, exemplified by WSIs sourced from a novel hospital. To address this challenge, this paper introduces the Retrieval-Augmented MIL (RAM-MIL) framework, which integrates Optimal Transport (OT) as the distance metric for nearest neighbor retrieval. The development of the RAM-MIL framework is motivated by both a theoretical finding, which demonstrates the approximation error in attention-based MIL can be reduced by shrinking input dimension, and empirical evidence establishing a connection between input dimension and the feature merging operation. Empirical evaluations conducted on WSI classification demonstrate that the proposed RAM-MIL framework achieves state-of-the-art performance in both in-domain scenarios, where the training and retrieval data are in the same domain, and more crucially, in out-of-domain scenarios, where the (unlabeled) retrieval data originates from a different domain. Furthermore, the use of the transportation matrix derived from OT renders the retrieval results interpretable at the instance level, in contrast to the vanilla  $l_2$  distance, and allows for visualization for human experts.

## 1 Introduction

As the standard supervised learning paradigm, single instance learning has been the focus of machine learning research and its performance has been improved significantly since the advent of deep learning [1]. Nonetheless, a notable drawback of single instance learning lies in its reliance on a substantial volume of labeled data [2], which poses practical challenges as a result of the high expense and time-consuming nature of fine-grained data labeling. Consequently, multiple instance learning (MIL) [3] has gained increasing popularity, given its ability to be trained with limited supervision. This trend has particularly gained momentum within the domain of histopathology, where the analysis of medical images at a gigabyte scale, such as whole slide images (WSI), has emerged as a focal point in recent studies [4, 5, 6, 7, 8, 9, 10, 11].

Even though the performance of MIL on cancer diagnosis [13, 12] is remarkable, existing algorithms are only evaluated on the in-domain test set and have the risk of performance degradation when confronted with distribution shifts between the training and test sets [14, 15]. The distribution shift problem is particularly important in the context of automated medical diagnosis models, since the models are deployed across diverse hospitals in distinct regions, each may employing varying imaging techniques. The robustness of such models in the face of distribution shift is of utmost importance within this safety-critical application, as any failure to address this challenge may result in severe consequences.

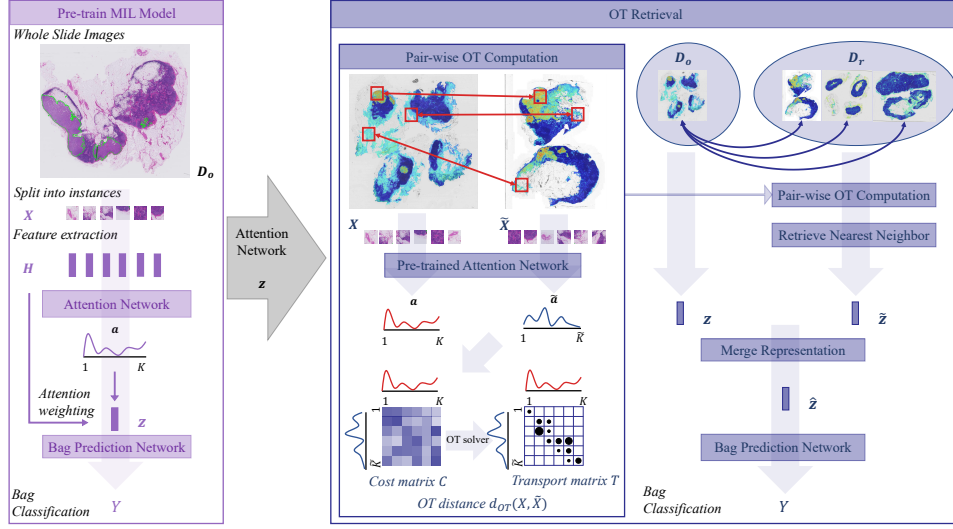


Figure 1: The proposed RAM-MIL method. **Left:** An attention-based MIL model is pre-trained on the training set to produce feature representations and attention weights. **Right:** With pre-trained features and attention weights, Optimal-Transport (OT) is employed to compute bag-to-bag distance, based on which the nearest neighbor bag from the retrieval set is selected. A training bag’s feature is then merged with its retrieved bag, which is used as the input feature to a bag classifier.

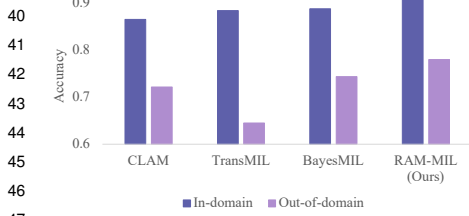


Figure 2: The comparisons of accuracy of MIL methods under in-domain and out-of-domain (OOD) settings. The out-of-domain is tested on Camelyon17 [12] with models trained on Camelyon16 [13]. MIL fails to generalize well to OOD data, while the proposed RAM-MIL improves the OOD performance as well as the in-domain performance.

calculate the distance between the two discrete distributions of two bags. Following retrieval, the original bags are merged with their retrieved counterparts from a retrieval dataset, employing a merge function such as convex combination, to generate feature representations. Subsequently, a classifier is trained as the MIL model. The retrieval dataset can either be the original dataset in the absence of out-of-domain data, or an *unlabeled* out-of-domain dataset in the out-of-domain adaptation setting. Our empirical study confirms that RAM-MIL effectively reduces the intrinsic dimensionality of the input feature space, surpassing existing MIL methods in both in-domain and out-of-domain settings. The advantage of OT distance over  $l_2$  and OT-variant distances is demonstrated by our ablation study. In addition, the OT-based retrieval result can be visualized using the transportation matrix, which makes the retrieval process interpretable. Our contributions are three-fold:

1. This work is among the first to investigate the out-of-domain performance for MIL, which is vital issue for the application of MIL in automated medical diagnosis with WSI. Our empirical study exposes the risk of recent MIL algorithms under distribution shifts.
2. Our theoretical result based on Wasserstein distance in the input space shows a negative correlation between MIL performance and the input data dimension, which motivates us to propose the novel RAM-MIL framework based on the OT as the distance metric in the retrieval process.
3. Our experiment demonstrates that RAM-MIL outperforms state-of-the-art MIL methods in both in-domain and out-of-domain adaptation scenarios. And the transportation matrix of OT distance provides a tool to visualize the correspondence between original and retrieval instances and makes our retrieval method interpretable.

## 2 Motivation of RAM-MIL

In this section, we show that why using OT could lead to a performance improvement for multiple instance learning.

We follow the standard formulation of Attention-based Multiple Instance Learning (MIL) [3, 4]. In MIL, the input is a bag of instances,  $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_K\}$ ,  $\mathbf{x}_k \in \mathbb{R}^D$ ,  $\mathbf{X} \in \mathcal{X}$  and  $K$  is the number of instances, which varies for different bags. There is a bag-level label  $Y$ , which is positive if at least one of the instances is positive, and negative if all instances are negative. We further assume the instances also have corresponding instance-level labels  $\{y_1, \dots, y_K\}$ , which are *unknown* during training. There are  $N$  such bag-label pairs constituting the dataset  $\mathcal{D} = \{\mathbf{X}_n, Y_n\}_{n=1}^N$ . The objective of MIL is to learn an optimal function for predicting the bag-level label with the bag of instances as input. To this end, the MIL model should be able to aggregate the information of instances  $\{\mathbf{x}_k\}_{k=1}^K$  to make the final decision. A well-adopted aggregation method is the embedding-based approach which maps  $\mathbf{X}$  to a bag-level representation  $\mathbf{z} \in \mathbb{R}^d$  and use  $\mathbf{z}$  to predict  $Y$ . [3] extends the embedding-based aggregation approach by leveraging the attention mechanism, namely attention-based deep MIL (ABMIL). First, a transformation  $g(\cdot)$  computes a low-dimensional embedding  $\mathbf{h}_k = g(\mathbf{x}_k) \in \mathbb{R}^d$  for each instance  $\mathbf{x}_k$ . The attention module aggregates the set of embeddings  $\{\mathbf{h}_k\}_{k=1}^K$  into a bag level embedding  $\mathbf{z} = \sum_{k=1}^K a_k \mathbf{h}_k = \sum_{k=1}^K a_k g(\mathbf{x}_k) = \mathcal{A}(\mathbf{X})$ , where the attention  $a_k$  for the  $k$ -th instance is computed via a softmax function,

$$a_k = \frac{\exp\{\mathbf{m}^T (\tanh(\mathbf{V}_1^T \mathbf{h}_k) \odot \text{sigmoid}(\mathbf{V}_2^T \mathbf{h}_k))\}}{\sum_{j=1}^K \exp\{\mathbf{m}^T (\tanh(\mathbf{V}_1^T \mathbf{h}_j) \odot \text{sigmoid}(\mathbf{V}_2^T \mathbf{h}_j))\}}. \quad (1)$$

where  $\mathbf{m}$ ,  $\mathbf{V}_1$ ,  $\mathbf{V}_2$  are the network parameters. The bag embedding  $\mathbf{z}$  is then mapped to the logits  $\mathbf{u}$  with a feed forward layer with parameter  $\mathbf{W}$  for the bag-level classification,  $\mathbf{u} = \mathbf{W}^T \mathbf{z}$ .

Assuming there is a ground-truth set scoring function  $S(\mathbf{X}) : \mathcal{X} \mapsto \mathbb{R}$  that generates the label  $Y$ , our task is to approximate the function with the attention-based model. We have the following theorem demonstrating that, when employing the Wasserstein distance in the input space, the approximation error of a Lipschitz continuous set scoring function is upper bounded by a term associated with the input dimension  $d$ . The insight from the theorem offers a guidepost for the design of our methodology.

**Theorem 1.** *Suppose  $S(\cdot)$  is a Lipschitz continuous function with respect to Wasserstein distance  $\mathcal{W}_p$  ( $1 \leq p < \infty$ ) and the Lipschitz constant is  $L$ . The bag of instances  $\mathbf{X}$  is sampled from a probability distribution  $\mu(\mathbf{x})$  with distribution dimension  $d_\mu$  (intrinsic dimension). For any invertible map  $\Phi : \mathcal{X} \rightarrow \mathbb{R}^d$ ,  $\exists$  function  $\sigma$  and  $\gamma$ , such that for any  $\mathbf{X} \in \mathcal{X}$ ,*

$$|S(\mathbf{X}) - \gamma(\Phi_{\mathbf{X} \in \mathcal{X}}\{\sigma(\mathbf{x}) : \mathbf{x} \in \mathbf{X}\})| \leq O(L \cdot K^{-\frac{1}{d_\mu}}). \quad (2)$$

The theorem has two implications for the ABMIL problem. First, using the Wasserstein metric in the input space, we establish this dimension-dependent error bound for the approximation, indicating the good property of Wasserstein metric in the theoretical sense. Second, this dimension-dependent error bound motivates our RAM-MIL framework to reduce the intrinsic dimensionality of the input space for improving the performance of ABMIL, which is consistent with a previous work [16] that shows feature mixup is effective at reducing input feature dimension. This paper uses OT distance as the instantiation of the Wasserstein distance, since OT with Euclidean distance as the cost is equivalent to  $\mathcal{W}_1$ . We use the entropy-regularized OT as the approximation of OT distance in our implementation since the regularization makes the optimization process more stable and efficient and is widely used in machine learning and vision research [17, 18]. The incorporation of OT into Multiple Instance Learning (MIL) provides a new framework for understanding and assessing the similarity between different instances. Intuitively, each bag in MIL can be considered as a probability distribution of instances, where the probabilities are the attention values  $\mathbf{a}$ . Thus, OT offers a rigorous and meaningful metric to quantify the similarity between two distributions. It takes into account not only the individual differences of instances but also their global arrangement within a bag, hence, providing a holistic comparison. Note that although Theorem 1 is proved for the input space  $\mathcal{X}$ , it is trivial to prove a similar result for the feature space  $\mathcal{H}$  (see the supplemental).

Building upon the benefits brought by OT, we propose a retrieval mechanisms to further strengthen the MIL framework. The key intuition is to leverage the rich information available across different bags. This allows for more robust and generalizable learning as each bag might not exist in isolation but is part of a broader set of data with shared and contrasting characteristics. The retrieval process, particularly when guided by OT, helps to identify and bring into context ‘‘neighbor’’ bags that are

---

**Algorithm 1** Retrieval-Augmented Multiple Instance Learning (RAM-MIL) Algorithm

---

- 1: Pre-train an MIL model on  $\mathcal{D}_o = \{\mathbf{X}_n, Y_n\}_{n=1}^{N_o}$ .
  - 2: Extract the sets of instance representations, i.e.,  $\mathcal{H}_o = \{\mathbf{H}_n\}_{n=1}^{N_o}$  and  $\mathcal{H}_r = \{\tilde{\mathbf{H}}_m\}_{m=1}^{N_r}$ , and compute their attentions, i.e.,  $\{\mathbf{a}_n\}_{n=1}^{N_o}$  and  $\{\tilde{\mathbf{a}}_m\}_{m=1}^{N_r}$ .
  - 3: Extract the sets of bag representations  $\{\mathbf{z}_n\}_{n=1}^{N_o}$  and  $\{\tilde{\mathbf{z}}_m\}_{m=1}^{N_r}$  with pre-trained model.
  - 4: **Input:**  $\{\mathbf{X}_n, Y_n, \mathbf{H}_n, \mathbf{z}_n, \mathbf{a}_n\}_{n=1}^{N_o}, \{\tilde{\mathbf{X}}_m, \tilde{Y}_m, \tilde{\mathbf{H}}_m, \tilde{\mathbf{z}}_m, \tilde{\mathbf{a}}_m\}_{m=1}^{N_r}$
  - 5: **for**  $n$  from 1 to  $N_o$  **do**
  - 6:      $\nu^* = \arg \min_{\tilde{\mathbf{H}}_m \in \mathcal{H}_r} d_{OT}(\mu_n, \nu_m)$ , where  $d_{OT}(\mu_n, \nu_m)$  is solved by (4)
  - 7:     Store  $\hat{\mathbf{z}}_n = \pi(\mathbf{z}_n, \mathbf{z}_{m^*})$ , where  $m^* = \text{getIndex}(\nu^*)$
  - 8: **end for**
  - 9: Train a logistic regression bag classifier on  $\{\hat{\mathbf{z}}_n, Y_n\}_{n=1}^{N_o}$ .
- 

most relevant or similar to a given bag. This allows the model to learn more effectively by taking into account the broader context in which each bag exists. Empirically, we confirm the OT-guided retrieval could effectively reduce the intrinsic dimension of input feature space, which explains the performance of proposed methodology. Moreover, with the computed transport matrix, the correspondence between the original and selected bag can be visualized, which makes the retrieval and classification process more interpretable/explainable.

### 3 Methodology

This section presents the RAM-MIL algorithm based on OT distance and its application in the out-of-domain adaptation.

#### 3.1 Retrieval-Augmented Multiple Instance Learning based on Optimal Transport

We introduce the methodology that utilizes OT distance to address the retrieval problem for multiple instance learning (MIL). The idea of enhancing MIL using retrieval lies in augmenting a bag feature from the original dataset with its nearest neighbor bag feature from the retrieval dataset. Consider the original dataset of bags marked for model training and validation, denoted as  $\mathcal{D}_o = \{\mathbf{X}_n, Y_n\}_{n=1}^{N_o}$ , and the unlabeled retrieval set represented as  $\mathcal{D}_r = \{\tilde{\mathbf{X}}_m\}_{m=1}^{N_r}$ , where  $N_o$  and  $N_r$  signify the number of bags in the two respective sets. For each instance  $\mathbf{x}_k$  in a bag  $\mathbf{X}$ , we extract an instance representation  $\mathbf{h}_k = g(\mathbf{x}_k)$ , where  $g(\cdot)$  is a neural net for feature extraction, such as ResNet50 [19]. Recall that an MIL model maps the bag of instance representations to the logits  $\mathbf{u}$  for loss computation, expressed as  $\mathbf{u} = \mathbf{W}^T \mathbf{z}$ . The intermediate representation  $\mathbf{z} = \mathcal{A}(\mathbf{X})$ , computed by averaging instance features and attention weights  $\mathbf{z} = \mathcal{A}(\mathbf{X}) = \sum_{k=1}^K a_k \mathbf{h}_k = \sum_{k=1}^K a_k g(\mathbf{x}_k)$ , is regarded as the *bag representation*.

The objective of the retrieval scheme is to select the bag in the retrieval set that is nearest to the input bag. To perform bag-to-bag retrieval, we use the distributional distance between two bags, which considers their constituent instances. Once the retrieved bag is obtained, its bag representation  $\tilde{\mathbf{z}}^*$  is then merged with the original bag representation  $\mathbf{z}$ . The merged feature, denoted as  $\hat{\mathbf{z}} = \pi(\mathbf{z}, \tilde{\mathbf{z}}^*)$ , is used for bag classification, where  $\pi(\cdot, \cdot)$  is the merging function, e.g., a convex combination.

**Retrieval with Optimal Transport.** To leverage the probabilistic geometry between instances, instead of directly retrieving the nearest bags based on the bag representations  $\mathbf{z}$ , our retrieval operates on the bags of instance representations  $\mathcal{H}_r = \{\tilde{\mathbf{H}}_m\}_{m=1}^{N_r} = \{\{\tilde{\mathbf{h}}_{mk}\}_{k=1}^K\}_{m=1}^{N_r}$ . Here  $K$  represents the number of instances,  $\tilde{\mathbf{H}}_m$  indicates the *bag of instance representations* and  $\tilde{\mathbf{h}}_{mk}$  is an *instance representation* derived via  $\tilde{\mathbf{h}}_{mk} = g(\tilde{\mathbf{x}}_{mk})$ . For each bag in  $\mathcal{D}_o$ , we similarly extract bags of instance representations as  $\mathcal{H}_o$ .

For each bag  $\mathbf{H} \in \mathcal{H}_o$ , we find its nearest neighbor in  $\mathcal{H}_r$  using the following OT problem to calculate distance between the two bags  $\mathbf{H}$  and  $\tilde{\mathbf{H}} \in \mathcal{H}_r$ . Let  $\mu = \sum_{i=1}^K a_i \delta_{\mathbf{h}_i}$  and  $\nu = \sum_{j=1}^{\tilde{K}} \tilde{a}_j \delta_{\tilde{\mathbf{h}}_j}$  be discrete distributions supported on  $\mathbf{H} = \{\mathbf{h}_k\}_{k=1}^K$  and  $\tilde{\mathbf{H}} = \{\tilde{\mathbf{h}}_k\}_{k=1}^{\tilde{K}}$ , where  $\delta_{\mu} = \delta(\mathbf{h} - \mu)$  is the translated Dirac delta function. Here,  $a_i$  and  $\tilde{a}_j$  are the attention scores from the MIL model for instance  $i$  in bag  $\mathbf{H}$  and instance  $j$  in bag  $\tilde{\mathbf{H}}$ , respectively, forming the probability vectors for the bags. The OT problem between these two distributions can be formulated as follows:

$$d_{OT}(\mu, \nu) = \min_{T \in \mathcal{T}(\mathbf{a}, \tilde{\mathbf{a}})} \sum_{i=1}^K \sum_{j=1}^{\tilde{K}} c(\mathbf{h}_i, \tilde{\mathbf{h}}_j) T_{ij} \quad (3)$$

In this equation,  $T$  denotes the transport plan matrix where each element  $T_{ij}$  specifies the amount of "mass" to be transported from  $\mathbf{h}_i$  to  $\tilde{\mathbf{h}}_j$ . The function  $c(\mathbf{h}_i, \tilde{\mathbf{h}}_j)$  is a cost function that quantifies the cost of transporting a unit of mass from  $\mathbf{h}_i$  to  $\tilde{\mathbf{h}}_j$ . A common choice of  $c(\mathbf{h}_i, \tilde{\mathbf{h}}_j)$  is the squared  $l_2$  distance between the instance features, i.e.,  $c(\mathbf{h}_i, \tilde{\mathbf{h}}_j) = \|\mathbf{h}_i - \tilde{\mathbf{h}}_j\|_2^2$ .  $\mathcal{T}(\mathbf{a}, \tilde{\mathbf{a}})$  represents the set of all matrices  $T$  that satisfy the marginals  $T^\top \mathbf{1}_K = \mathbf{a}$  and  $T \mathbf{1}_{\tilde{K}} = \tilde{\mathbf{a}}$ . Here,  $\mathbf{1}_n$  and  $\mathbf{1}_m$  are vectors of ones. We also impose an entropy regularization [20] term to reduce the sensitivity to outlier instances. Specifically, we solve the following OT problem with a Sinkhorn's algorithm [20]:

$$d_{OT}(\mu, \nu) = \min_{T \in \mathcal{T}(\mathbf{a}, \tilde{\mathbf{a}})} \sum_{i=1}^K \sum_{j=1}^{\tilde{K}} c(\mathbf{h}_i, \tilde{\mathbf{h}}_j) T_{ij} + \beta \cdot \sum_{ij} T_{ij} \log T_{ij} \quad (4)$$

$$s.t., T^\top \mathbf{1}_K = \mathbf{a}, T \mathbf{1}_{\tilde{K}} = \tilde{\mathbf{a}}, T \geq 0$$

In the context of the retrieval-based MIL problem, the nearest neighbor bag for a given bag  $\mathbf{H}$  would be the one  $\tilde{\mathbf{H}}$  in the retrieval set that results in the smallest OT cost with  $\mathbf{H}$ :

$$\nu^* = \arg \min_{\tilde{\mathbf{H}} \in \mathcal{H}_r} d_{OT}(\mu, \nu_m) \quad (5)$$

In the retrieval process, we use an MIL model pre-trained on  $\mathcal{D}_o$  to extract the attention vectors  $\mathbf{a}$  and  $\tilde{\mathbf{a}}$ , which serve as mass probability distributions for calculating the OT, in addition to extracting bag representations  $\mathbf{z}$  and  $\tilde{\mathbf{z}}$ . This method effectively transforms the instance importance, as determined by the model, into a mass probability distribution. Instances with high attention scores are considered to possess larger "mass". In the context of the OT problem, such instances become more "expensive" to transport, thereby significantly influencing the computation of the transport cost. Consequently, the model is encouraged to pair similar instances carrying substantial mass, a strategy that aligns with the intuitive notion of matching similar instances together. Employing the pre-extracted attention to classify the weighted bag representation after retrieval ensures consistency with the retrieval process and enhances efficiency. The whole retrieval-enhanced multiple instance learning is shown in Algo. 1.

### 3.2 Unsupervised Domain Adaptation

The proposed methodology can be effectively employed in the context of unsupervised domain adaptation (UDA) [21]. UDA seeks to leverage labeled data from a source domain to train a model that can perform well on an unlabeled target domain.

Remarkably, our retrieval algorithm (as described in Algo. 1) can be seamlessly applied to the UDA task. To achieve this, we set  $\mathcal{H}_o$  to represent the source domain and  $\mathcal{H}_r$  to represent the target domain. Our approach retrieves the nearest neighbor bag representation from the target domain for each bag representation in the source domain. Subsequently, we train the final classifier using the merged representation and the source domain label.

The RAM-MIL based on OT allows leveraging the geometric structure of the instance space across domains. This promotes the alignment of the source and target domains at the instance level. By merging the bag features, the discriminative information embedded in the source domain is transferred to the target domain. As a result, the bag classifier is able to discriminate the target domain bags without needing access to their labels.

### 3.3 Reduction of Intrinsic Dimension

In this section, we validate our hypothesis that retrieval can effectively reduce the intrinsic dimension of each bag's distribution. We use the two NN estimator [22] for computing the intrinsic dimension for the bag representations, either the original features for MIL or merged features for retrieval models. Figure 3 shows the intrinsic dimension of CLAM [4], Manifold Mixup [16], OT based retrieval and different variants of retrieval methods. The  $l_2$  distance based retrieval, approximate

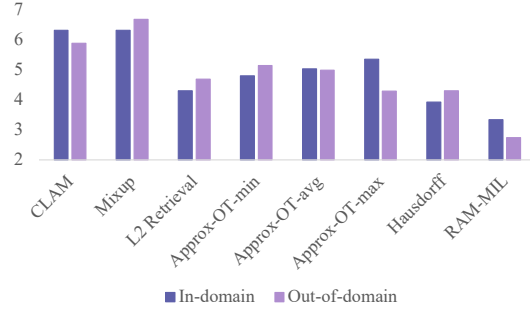


Figure 3: The intrinsic dimension of bag representations for different MIL methods and retrieval models. CLAM uses the original feature and the rest use merged feature from retrieval. RAM-MIL is the most effective method at reducing intrinsic dimension of feature space.

OT retrieval and Hausdorff distance based retrieval are variants of our retrieval method that will be elaborated in Section 5.

The figure shows that our proposed retrieval methods indeed lower the intrinsic dimensions of bag representations in both in-domain and out-of-domain settings. As shown in (2), the approximation performance of a ground-truth set scoring function is theoretically constrained by the intrinsic dimension of the distribution, when measured under the Wasserstein distance. RAM-MIL delivers the lowest intrinsic dimensions among the methods compared, thus validating its superior performance in classification and aligns with our theoretical analysis (2). We offer more detailed discussion of intrinsic dimension in Section 5.2.

## 4 Related works

**Attention-based multiple instance learning** ABMIL [3] first introduced the attention mechanism for MIL to an interpretation of importance for instances. Building on this, DSMIL [9] used contrastive learning for feature extraction and established global connections between instance attentions. TransMIL [5] took this a step further and developed a correlated MIL, employing multi-head self-attention and spatial information encoding for thorough global correlation. CLAM [4] extends ABMIL to the case of multiple classes and builds an integrated toolbox for visualizing the uncertainty. DTFD-MIL [23] proposes a two-stage feature distillation MIL framework for enhancing the performance. Bayes-MIL [11] studies the fundamental interpretability problem in the attention based MIL framework and proposed to address it with a probabilistic method. In our RAM-MIL, the attention weight is utilized as a measure of the probability density, indicating the amount of “mass” being moved from the source bag to the target bag. Greater attention weights in RAM-MIL indicate more significant relationships between instances in the bags being compared.

The preceding studies on whole slide images domain generalization have primarily focused on instance-level classification [24, 25, 26, 27, 28], rather than tackling the more complex challenge of weakly-supervised bag-level classification (MIL). In contrast, our approach seeks to address this more complex problem of unsupervised domain adaptation for MIL. Therefore, we have chosen not to include those works in this discussion due to the significantly different focus of their methodology.

## 5 Experiments

The proposed methodology is evaluated on whole slide image (WSI) datasets (Camelyon16 and Camelyon17 [29, 13, 12]) and general MIL datasets (See results in supplemental). The evaluation is divided into in-domain and UDA settings. For the in-domain setting, the retrieval set  $\mathcal{H}_r$  is the union of training and validation set of  $\mathcal{H}_o$ . For the UDA setting, the retrieval set is a held-out dataset with no labels involved in training or inference.

We compare our OT-based retrieval with Euclidean distance retrieval, Hausdorff distance retrieval and Manifold MixUp as baselines. The Euclidean ( $l_2$  norm) distance  $d_{l_2}(\cdot, \cdot)$  and Hausdorff distance based retrieval are implemented in the following ways:

- **Direct bag representation retrieval.** As a baseline method, we propose an approach to augment bag features by directly retrieving the nearest neighbor. The setup follows the initial steps of Algo. 1, from line 1 to line 3. Instead of computing the instance correlations, we directly compute the  $l_2$  distance between bag representations, formulated as  $d_{l_2}(\mathbf{H}_n, \tilde{\mathbf{H}}_m) = \|\mathbf{z}_n - \tilde{\mathbf{z}}_m\|_2^2$ . For the  $n$ -th bag representation, we select the nearest neighbor from the retrieval set:  $\tilde{\mathbf{H}}_{m^*} = \arg \min_{\tilde{\mathbf{H}}_m \in \mathcal{H}_r} d_{l_2}(\mathbf{H}_n, \tilde{\mathbf{H}}_m)$ . The merged bag representation  $\hat{\mathbf{z}}_n = \pi(\mathbf{z}_n, \mathbf{z}_{m^*})$  is subsequently employed for bag classification, where  $m^* = \text{getIndex}(\tilde{\mathbf{H}}_{m^*})$ .
- **Approximate OT retrieval.** Recall that prior to applying the Sinkhorn’s algorithm to solve OT, it is necessary to compute the cost matrix  $C = [c(\mathbf{h}_i, \tilde{\mathbf{h}}_j)]_{ij} = \|\mathbf{h}_i - \tilde{\mathbf{h}}_j\|_2^2$ , where each element is the squared  $l_2$  distance between instance representations. To highlight the significance of solving the OT problem, we introduce three simple approximations of OT by directly employing the minimum, average and maximum values of the cost matrix. The  $d_{\text{Cmin}}(\mathbf{H}_n, \tilde{\mathbf{H}}_m) = \min_{ij} [c(\mathbf{h}_i, \tilde{\mathbf{h}}_j)]_{ij}$  focuses on the closest and most similar instances between bags. The  $d_{\text{Cavg}}(\mathbf{H}_n, \tilde{\mathbf{H}}_m) = \text{average}_{ij} [c(\mathbf{h}_i, \tilde{\mathbf{h}}_j)]_{ij}$  offers a holistic measure of bag similarity but might lose some finer details. The  $d_{\text{Cmax}}(\mathbf{H}_n, \tilde{\mathbf{H}}_m) = \max_{ij} [c(\mathbf{h}_i, \tilde{\mathbf{h}}_j)]_{ij}$  focuses on the

Table 1: Results on CAMELYON16 and CAMELYON17 for in-domain classification and unsupervised domain adaptation, for comparisons with related methods.

|                            | In-Domain (CAM16)   |                     | Out-of-Domain (CAM17) |                     |
|----------------------------|---------------------|---------------------|-----------------------|---------------------|
|                            | AUC                 | Accuracy            | AUC                   | Accuracy            |
| DSMIL                      | 0.8944±0.051        | 0.8682±0.060        | -                     | -                   |
| CLAM                       | 0.9177±0.044        | 0.8650±0.060        | 0.7613±0.054          | 0.7214±0.044        |
| TransMIL                   | 0.9307±0.024        | 0.8837±0.041        | 0.5697±0.118          | 0.6451±0.100        |
| Bayes-MIL                  | 0.9432±0.049        | 0.8875±0.052        | 0.7839±0.044          | 0.7435±0.058        |
| Mixup Retr <sub>I</sub>    | 0.9260±0.051        | 0.8825±0.051        | -                     | -                   |
| Mixup Retr <sub>IO</sub>   | 0.9271±0.048        | 0.8850±0.046        | 0.7658±0.052          | 0.7594±0.044        |
| Mixup Retr <sub>O</sub>    | 0.9271±0.045        | 0.8825±0.048        | 0.7641±0.056          | 0.7593±0.058        |
| $l_2$ Retr <sub>I</sub>    | 0.9281±0.047        | 0.8800±0.055        | -                     | -                   |
| $l_2$ Retr <sub>IO</sub>   | 0.9241±0.048        | 0.8950±0.048        | 0.7627±0.055          | 0.7353±0.051        |
| $l_2$ Retr <sub>O</sub>    | 0.9398±0.045        | 0.8950±0.055        | 0.7738±0.050          | 0.7493±0.050        |
| RAM-MIL Retr <sub>I</sub>  | <b>0.9451±0.036</b> | 0.8925±0.050        | -                     | -                   |
| RAM-MIL Retr <sub>IO</sub> | 0.9365±0.052        | <b>0.9200±0.050</b> | <b>0.7974±0.054</b>   | 0.7433±0.073        |
| RAM-MIL Retr <sub>O</sub>  | 0.9419±0.048        | 0.9175±0.051        | 0.7681±0.058          | <b>0.7795±0.021</b> |

most dissimilar instances between the two bags. These three measures can be considered as OT approximations that use information from the squared  $l_2$  cost matrix.

- **Hausdorff distance retrieval.** The symmetric Hausdorff distance is employed for retrieval, computed as  $d_H(\mathbf{H}_n, \mathbf{H}_m) = \max\{d_h(\mathbf{H}_n, \mathbf{H}_m), d_h(\mathbf{H}_m, \mathbf{H}_n)\}$ . Here,  $d_h(\mathbf{H}_n, \mathbf{H}_m) = \max_{\mathbf{h}_i \in \mathbf{H}_n} \min_{\tilde{\mathbf{h}}_j \in \tilde{\mathbf{H}}_m} \|\mathbf{h}_i - \tilde{\mathbf{h}}_j\|_2^2$  is the directed Hausdorff distance. The Hausdorff distance is an appropriate retrieval metric when assessing the maximum dissimilarity between two sets, emphasizing the instances that differ the most between the sets. In comparison, OT provides a more comprehensive and versatile choice for retrieval tasks, given its ability to consider all possible matchings between elements in the sets and finds an optimal one.

## 5.1 Classification of Whole Slide Image

**Experimental setup:** In this section, we validate the methodology within both in-domain and out-of-domain contexts. In both contexts, the training set is from Camelyon16. In the in-domain setting, we evaluate our models using the Camelyon16 dataset, where the combination of training and validation sets is treated as  $\mathcal{D}_r$ , which is denoted as  $\text{Retr}_I$ . For the out-of-domain context, we use Camelyon16 as  $\mathcal{D}_o$ . As for  $\mathcal{D}_r$ , we consider two configurations: 1) retrieval from the training and validation sets of both Camelyon16 and Camelyon17, and 2) retrieval solely from the training and validation sets of Camelyon17, referred to as  $\text{Retr}_{IO}$  and  $\text{Retr}_O$ , respectively. To ensure robustness and avoid the influence of outliers, each experiment is executed 10 times on randomly partitioned train/validation/test sets.

The baseline model we use for comparison is CLAM, an extension of attention-based Multiple Instance Learning (MIL) that incorporates an instance clustering loss. In our case, the pre-trained model is a CLAM model for binary classification. We also make comparisons with other models such as DSMIL, TransMIL, Bayes-MIL, and Mixup. For the ablation study, we add our proposed  $l_2$  retrieval method with direct bag representation retrieval. For both the Mixup and  $l_2$  retrieval models, we ensure that the source of the augmented data aligns with our in-domain and out-of-domain settings. Note that Mixup and our proposed methods use CLAM as the MIL method. Furthermore, Mixup shares the same merging function with our retrieval method, which is a simple addition or a convex combination, in the reported experiments. Our method does not need the ground-truth label information for retrieval set while Mixup uses that for the in-domain experiments. More results about the merging function is shown in the supplemental.

When we compute the RAM-MIL, we often deal with very large WSIs that can contain over 100,000 instances. This makes the calculation of OT quite time-consuming. To speed things up, we first perform pre-selection of potentially important instances. Instead of using all the instances in a bag  $\mathcal{H}_n$ , we pick the top 20% of instances based on their attention scores and normalized the attention scores (Ablation study for the percentage is in supplemental). We provide more detailed comparisons in percentage of instances used, in the supplemental. In this way, we are focusing on the most important instances and making the whole process more efficient.

Table 2: Ablation study evaluated on CAMELYON16 and CAMELYON17.

|                                  | In-Domain (CAM16)   |                     | Out-of-Domain (CAM17) |                     |
|----------------------------------|---------------------|---------------------|-----------------------|---------------------|
|                                  | AUC                 | Accuracy            | AUC                   | Accuracy            |
| Approx-OT-min Retr <sub>I</sub>  | 0.9301±0.033        | 0.8800±0.043        | -                     | -                   |
| Approx-OT-min Retr <sub>IO</sub> | 0.8601±0.080        | 0.8000±0.062        | 0.7640±0.058          | 0.7353±0.043        |
| Approx-OT-min Retr <sub>O</sub>  | 0.9196±0.063        | 0.8750±0.056        | 0.7680±0.065          | 0.7274±0.046        |
| Approx-OT-avg Retr <sub>I</sub>  | 0.9227±0.038        | 0.8600±0.036        | -                     | -                   |
| Approx-OT-avg Retr <sub>IO</sub> | 0.8865±0.063        | 0.8425±0.061        | 0.7625±0.060          | 0.7174±0.041        |
| Approx-OT-avg Retr <sub>O</sub>  | 0.9195±0.063        | 0.8750±0.056        | 0.7641±0.054          | 0.7374±0.061        |
| Approx-OT-max Retr <sub>I</sub>  | 0.9174±0.049        | 0.8625±0.044        | -                     | -                   |
| Approx-OT-max Retr <sub>IO</sub> | 0.9182±0.047        | 0.8650±0.048        | 0.7576±0.060          | 0.7214±0.062        |
| Approx-OT-max Retr <sub>O</sub>  | 0.9195±0.063        | 0.8750±0.056        | 0.7681±0.065          | 0.7414±0.038        |
| $l_2$ Retr <sub>I</sub>          | 0.9281±0.047        | 0.8800±0.055        | -                     | -                   |
| $l_2$ Retr <sub>IO</sub>         | 0.9241±0.048        | 0.8950±0.048        | 0.7627±0.055          | 0.7353±0.051        |
| $l_2$ Retr <sub>O</sub>          | 0.9398±0.045        | 0.8950±0.055        | 0.7738±0.050          | 0.7493±0.050        |
| Hausdorff Retr <sub>I</sub>      | 0.9273±0.057        | 0.8850±0.049        | -                     | -                   |
| Hausdorff Retr <sub>IO</sub>     | 0.9226±0.055        | 0.8725±0.060        | 0.7651±0.054          | 0.7434±0.051        |
| Hausdorff Retr <sub>O</sub>      | 0.9322±0.046        | 0.8950±0.048        | 0.7695±0.056          | 0.7513±0.051        |
| RAM-MIL Retr <sub>I</sub>        | <b>0.9451±0.036</b> | 0.8925±0.050        | -                     | -                   |
| RAM-MIL Retr <sub>IO</sub>       | 0.9365±0.052        | <b>0.9200±0.050</b> | <b>0.7974±0.054</b>   | 0.7433±0.073        |
| RAM-MIL Retr <sub>O</sub>        | 0.9419±0.048        | 0.9175±0.051        | 0.7681±0.058          | <b>0.7795±0.021</b> |

**In-domain comparisons:** Table 1 presents the in-domain results. Mixup,  $l_2$  retrieval, and RAM-MIL all enhance the classification performance for in-domain data beyond the CLAM baseline. The straightforward  $l_2$  retrieval-based MIL outpaces Mixup, demonstrating that selecting the nearest neighbor using a basic metric offers superior augmentation compared to random data selection. Using OT as a distance metric, the retrieval-based MIL method surpasses all the other compared methods for the in-domain dataset. Notably, the Retr<sub>IO</sub> configuration, which employs the most extensive data set for retrieval (CAMELYON16 and CAMELYON17), offers the most significant boost in in-domain classification accuracy.

**Out-of-domain comparisons:** The right-hand side of Table 1 showcases the results for the out-of-domain setting. Notably, all retrieval methods, including ours, enhance the classification performance for out-of-domain data, without accessing its labels. This suggests that retrieval benefits from the transfer of representative information from the source to the target domain via the merging of bag features. The  $l_2$  retrieval, while trailing Mixup in terms of accuracy, surpasses it in terms of AUC. Our OT-based retrieval achieves the highest performance across both AUC and accuracy metrics. The underlying reason for the good performance might be it could consider all possible matchings between elements in the sets and finds an optimal one.

**Ablation study** Table 2 presents the ablation study using different retrieval methods, which shows that even  $l_2$  retrieval, which operates without considering instance-level distances, surpasses the performance of the three OT approximation methods in both in-domain and out-of-domain settings. Among these, only the Approx-OT-min, when used for in-domain data retrieval, achieves a performance comparable to that of  $l_2$  retrieval. This underscores the fact that relying on simple statistics of the cost matrix cannot provide a robust and consistent approximation of the optimal transport. This outcome reinforces the necessity of solving the complete optimal transport problem, using Sinkhorn’s algorithm, to attain reliable results.

## 5.2 Dimensionality Reduction

We next show that retrieval-based methods can effectively reduce the intrinsic dimension. Figure 3 shows the intrinsic dimension for various methods. The proposed RAM-MIL based on optimal transport delivers the lowest intrinsic dimension, among all methods, both retrieval and vanilla models. Interestingly, Mixup seems to increase the intrinsic dimension, suggesting that random selection and merging of bag representations do not necessarily lead to a more compact representation. This could explain why Mixup only offers marginal improvements.

Figure 3 shows that while the CLAM model has an intrinsic dimension between 5-7, RAM-MIL achieves a dimension between 2-4. Based on these observations, we choose dimension 5 to guide manual dimensionality reduction. The objectives of this exercise are two-fold: firstly, to validate



Table 3: Classification on CAMELYON16 with bag representations of reduced dimensions (5-dim).

|         | In-Domain (CAM16)                  |                                    | Out-of-Domain (CAM17)              |                                    |
|---------|------------------------------------|------------------------------------|------------------------------------|------------------------------------|
|         | AUC                                | Accuracy                           | AUC                                | Accuracy                           |
| CLAM    | 0.9216 $\pm$ 0.057                 | 0.6625 $\pm$ 0.177                 | 0.7486 $\pm$ 0.040                 | 0.4991 $\pm$ 0.105                 |
| RAM-MIL | <b>0.9273<math>\pm</math>0.072</b> | <b>0.8475<math>\pm</math>0.157</b> | <b>0.7509<math>\pm</math>0.074</b> | <b>0.7053<math>\pm</math>0.113</b> |

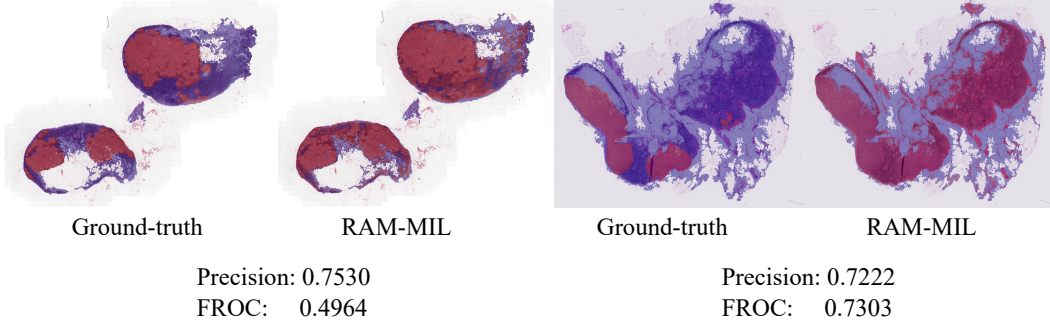


Figure 4: The visualization of predicted positive instances (red regions) in whole slide images for RAM-MIL. FROC is defined as the average sensitivity (recall) at 6 predefined false positive (FP) rates: 1/4, 1/2, 1, 2, 4 and 8 FPs. For reference, the averaged precision and FROC in Bayes-MIL [11] (one of the SOTA method designed for instance-level performance) are 0.8107 and 0.4919.

that intrinsic dimension can accurately explain performance, and secondly, to illustrate that simple methods like SVD are not necessarily beneficial for dimensionality reduction of bag representations. In our experiment, we employ SVD to reduce each bag representation of CLAM and each merged bag representation of RAM-MIL to a 5-dimensional vector. Then, we use these vectors and the bag labels to train a single logistic regression classifier. The results indicate that RAM-MIL significantly outperforms CLAM in classification accuracy. This experiment shows that a simple dimensionality reduction method like SVD does not significantly improve the performance of the CLAM baseline. However, RAM-MIL, which organizes the latent space by merging nearest neighbor features and then applies SVD, can achieve a commendable performance.

### 5.3 Interpretability with OT for MIL

We show that by using optimal transport retrieval, a novel method for highlighting the important instances could be derived. For a pair of source bag  $\mathbf{H}$  and target bag  $\tilde{\mathbf{H}}$ , solving the optimal transport problem generates the transport matrix  $\mathbf{T} = [T_{ij}]_{ij}$ . Suppose that we are interested in the instance labels of the source bag, and suppose that the instance labels of the target bag are known. The positive instances in the source bag are simply the instances transported to the positive instances in the target bag, i.e.,  $\hat{y}_i = \tilde{y}_{j^*} = \arg \max_j T_{ij}$ . Here,  $\hat{y}$  is the predicted instance label for the source bag, and  $\tilde{y}$  is the true instance label in the target bag. Figure 4 shows this method could provide a good coverage of the positive area, without model training or accessing the bag label. In contrast to OT, retrieval-based methods utilizing other distance metrics do not possess this specific property.

## 6 Conclusion

In this work, we investigate the out-of-domain performance of multiple instance learning model, which is crucial for the ML-assisted medical diagnosis with the whole slide images. An optimal transport based retrieval method is proposed based on the understanding that there exists a negative correlation between MIL performance and input data dimension. The proposed RAM-MIL framework works for both in-domain and out-of-domain, meanwhile outperforms state-of-the-art MIL methods. We hypothesize that OT-based retrieval could reduce the intrinsic dimension therefore improves the MIL performance and validate the hypothesis by computing the intrinsic dimension of bag representations. The transportation matrix of OT distance provides a tool to visualize the correspondence between original and retrieval instances and makes our retrieval method interpretable. A limitation is the efficiency of optimal transport solver is constrained by the number of instances within a bag. Although we have mitigated this problem by selecting partial instances based on high attention values, we anticipate that a more principled method will be sought in future works.

## References

- [1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.
- [2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- [3] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018.
- [4] Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6):555–570, 2021.
- [5] 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*, 34:2136–2147, 2021.
- [6] 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*, 35:29484–29497, 2022.
- [7] Linhao Qu, Manning Wang, Zhijian Song, et al. Bi-directional weakly supervised knowledge distillation for whole slide image classification. *Advances in Neural Information Processing Systems*, 35:15368–15381, 2022.
- [8] Syed Ashar Javed, Dinkar Juyal, Harshith Padigela, Amaro Taylor-Weiner, Limin Yu, and Aaditya Prakash. Additive mil: Intrinsically interpretable multiple instance learning for pathology. *Advances in Neural Information Processing Systems*, 35:20689–20702, 2022.
- [9] 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*, pages 14318–14328, 2021.
- [10] Jinxi Xiang and Jun Zhang. Exploring low-rank property in multiple instance learning for whole slide image classification. In *The Eleventh International Conference on Learning Representations*, 2023.
- [11] Yufei CUI, Ziquan Liu, Xiangyu Liu, Xue Liu, Cong Wang, Tei-Wei Kuo, Chun Jason Xue, and Antoni B. Chan. Bayes-MIL: A new probabilistic perspective on attention-based multiple instance learning for whole slide images. In *The Eleventh International Conference on Learning Representations*, 2023.
- [12] Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermesen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, et al. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. *IEEE transactions on medical imaging*, 38(2):550–560, 2018.
- [13] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, Meyke Hermesen, Quirine F Manson, Maschenka Balkenhol, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22):2199–2210, 2017.
- [14] Joaquin Quinonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. Mit Press, 2008.
- [15] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanus Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021.

- [16] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In *International conference on machine learning*, pages 6438–6447. PMLR, 2019.
- [17] Jia Wan, Ziquan Liu, and Antoni B Chan. A generalized loss function for crowd counting and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1974–1983, 2021.
- [18] Ziquan Liu, Yi Xu, Yuanhong Xu, Qi Qian, Hao Li, Xiangyang Ji, Antoni Chan, and Rong Jin. Improved fine-tuning by better leveraging pre-training data. *Advances in Neural Information Processing Systems*, 35:32568–32581, 2022.
- [19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [20] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.
- [21] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015.
- [22] Elena Facco, Maria d’Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific reports*, 7(1):12140, 2017.
- [23] 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*, pages 18802–18812, 2022.
- [24] Milad Sikaroudi, Maryam Hosseini, Ricardo Gonzalez, Shahryar Rahnamayan, and HR Tizhoosh. Generalization of vision pre-trained models for histopathology. *Scientific reports*, 13(1):6065, 2023.
- [25] Marc Aubreville, Nikolas Stathonikos, Christof A Bertram, Robert Klopffleisch, Natalie Ter Hoeve, Francesco Ciompi, Frauke Wilm, Christian Marzahl, Taryn A Donovan, Andreas Maier, et al. Mitosis domain generalization in histopathology images—the midog challenge. *Medical Image Analysis*, 84:102699, 2023.
- [26] Jiawei Yang, Hanbo Chen, Jiangpeng Yan, Xiaoyu Chen, and Jianhua Yao. Towards better understanding and better generalization of low-shot classification in histology images with contrastive learning. In *International Conference on Learning Representations*.
- [27] Jian Ren, Ilker Hacihaliloglu, Eric A Singer, David J Foran, and Xin Qi. Unsupervised domain adaptation for classification of histopathology whole-slide images. *Frontiers in bioengineering and biotechnology*, 7:102, 2019.
- [28] Jian Ren, Ilker Hacihaliloglu, Eric A Singer, David J Foran, and Xin Qi. Adversarial domain adaptation for classification of prostate histopathology whole-slide images. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16–20, 2018, Proceedings, Part II 11*, pages 201–209. Springer, 2018.
- [29] Geert Litjens, Peter Bandi, Babak Ehteshami Bejnordi, Oscar Geessink, Maschenka Balkenhol, Peter Bult, Altuna Halilovic, Meyke Hermesen, Rob van de Loo, Rob Vogels, et al. 1399 h&e-stained sentinel lymph node sections of breast cancer patients: the camelyon dataset. *GigaScience*, 7(6):giy065, 2018.
- [30] Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. *arXiv preprint arXiv:1707.00087*, 2017.
- [31] Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1-2):31–71, 1997.

- 467 [32] Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. Support vector machines for  
468 multiple-instance learning. *Advances in neural information processing systems*, 15, 2002.