Task-specific Fine-tuning via Variational Information Bottleneck for Weakly-supervised Pathology Whole Slide Image Classification

Honglin Li[1,3], Chenglu Zhu[2,3], Yunlong Zhang[1,3], Yuxuan Sun[1,3],
Zhongyi Shui[1,3], Wenwei Kuang[3,4], Sunyi Zheng[2,3], Lin Yang[2,3*]

[1]College of Computer Science and Technology, Zhejiang University
[2]Research Center for Industries of the Future and [3]School of Engineering, Westlake University
[4]The University of Hong Kong

{lihonglin,yanglin}@westlake.edu.cn

Abstract

While Multiple Instance Learning (MIL) has shown promising results in digital Pathology Whole Slide Image (WSI) analysis, such a paradigm still faces performance and generalization problems due to high computational costs and limited supervision of Gigapixel WSIs. To deal with the computation problem, previous methods utilize a frozen model pretrained from ImageNet to obtain representations, however, it may lose key information owing to the large domain gap and hinder the generalization ability without image-level training-time augmentation. Though Self-supervised Learning (SSL) proposes viable representation learning schemes, the downstream task-specific features via partial label tuning are not explored. To alleviate this problem, we propose an efficient WSI fine-tuning framework motivated by the Information Bottleneck theory. The theory enables the framework to find the minimal sufficient statistics of WSI, thus supporting us to fine-tune the backbone into a task-specific representation only depending on WSI-level weak labels. The WSI-MIL problem is further analyzed to theoretically deduce our fine-tuning method. We evaluate the method on five pathological WSI datasets on various WSI heads. The experimental results show significant improvements in both accuracy and generalization compared with previous works. Source code will be available at https://github.com/invoker-LL/WSI-finetuning.

1. Introduction

Digital Pathology or microscopic images have been widely used for the diagnosis of cancers such as Breast Cancer [13] and Prostate Cancer [6]. However, the reading of Whole Slide Images (WSIs) with gigapixel resolution is
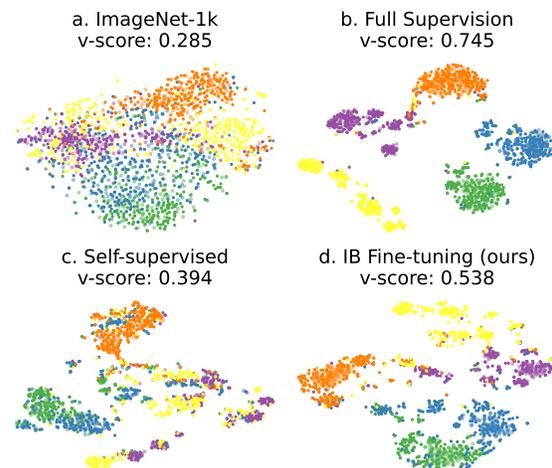
---
*Corresponding author.

Figure 1. T-SNE visualization of different representations on patches. Our method converts chaotic ImageNet-1K and SSL features into a more task-specific and separable distribution. The cluster evaluation measurement, v-scores, show weakly supervised fine-tuned features are more close to full supervision compared to others. a. ImageNet-1k pretraining. b. Full patch supervision. c. Self-supervised Learning. d. Fine-tuning with WSI labels.

time-consuming which poses an urgent need for automatic computer-assisted diagnosis. Though computers can boost the speed of the diagnosis process, the enormous size of resolution, over 100M [45], makes it infeasible to acquire precise and exhaustive annotations for model training, let alone the current hardware can hardly support the parallel training on all patches of a WSI. Hence, an annotation-efficient learning scheme with light computation is increasingly desirable to cope with those problems. In pathology WSI analysis, the heavy annotation cost is usually alleviated by Multiple Instance Learning (MIL) with only WSI-level weak supervision, which makes a comprehensive decision on a series of instances as a bag sample [19, 31]. Intuitively, all

small patches in the WSI are regarded as instances to constitute a bag sample [7, 30, 38], where the WSI's category corresponds to the max lesion level of all patch instances.

However, most methods pay much effort to design WSI architectures while overlooking the instance-level representation ability. Because of the computational limitation, the gradient at the WSI-level is impossible to parallelly back-propagate to instance encoders with more than 10k instances of a bag. Thus parameters of the pretrained back-bone from ImageNet-1k (IN-1K) are frozen to obtain invariant embeddings. Due to the large domain gap between IN-1K and pathological images, some essential information may be discarded by layers of frozen convolutional filters, which constrains the accuracy of previous WSI methods. To address the constraint, recent works [9, 25] make efforts to learn a good feature representation at the patch-level by leveraging Self-supervised Learning (SSL). However, such task-agnostic features are dominated by the proxy objective of SSL, e.g. Contrastive Learning in [8, 11, 17] may push away the distance between two instances within the same category, thus only performs slightly better than IN-1K pretraining in WSI classification. Nearly all SSL methods [8, 11, 16, 17] proposed on natural image recognition utilize a small portion of annotations to get promising fine-tuning accuracy compared to full supervision, which is higher than Linear Probing [17] by a large margin.

These findings illuminate us to design a fine-tuning scheme for WSI analysis to convert IN-K or SSL task-agnostic representations into task-specifics. Motivated by the Information Bottleneck (IB) theory [1, 2], we argue that pretraining is limited to downstream tasks, therefore fine-tuning is necessary for WSI analysis. In addition, we develop a solution based on Variational IB to tackle the dilemma of fine-tuning and computational limitation by its minimal sufficient statistics and attribution properties [1, 23]. The differences among the above three feature representations are depicted in Figure 1, where the feature representation under full patch-level supervision is considered as the upper bound.

Our main contributions are in 3 folds: 1) We propose a simple agent task of WSI-MIL by introducing an IB module that distills over 10k redundant instances within a bag into less than 1k of the most supported instances. Thus the parallel computation cost of gradient-based training on Gigapixel Images is over ten times relieved. By learning and making classification on the simplified bag, we find that there are trivial information losses due to the low-rank property of pathological WSI, and the distilled bag makes it possible to train a WSI-MIL model with the feature extractor on patches end-to-end, thus boosting the final performance. 2) We argue that the performance can be further improved by combining with the SSL pretraining since we could convert the task-agnostic representation from SSL into task-specific

one by well-designed fine-tuning. The proposed framework only relies on annotations at WSI levels, which is similar to recent SSL approaches [8, 11, 16, 17]. Note that our method only utilizes less than a 1% fraction of full patch annotation to achieve competitive accuracy compared to counterparts. 3) Versatile training-time augmentations can be incorporated with our proposed fine-tuning scheme, thus resulting in better generalization in various real-world or simulated datasets with domain shift, which previous works ignore to validate. These empirical results show that our method advances accuracy and generalization simultaneously, and thus would be more practical for real-world applications.

2. Related Work

2.1. Multiple Instance Learning for WSI Analysis

Multiple Instance Learning (MIL) is a well-defined task and has been explored extensively. Currently, there are mainly two lines of methods for WSI analysis: 1) explicit modeling the MIL definition $Y = pooling\{y_1, y_2, ..., y_n\}$ that WSI level prediction is aggregated by the probability of all patches with Mean or Max-pooling [7, 45]. 2) implicit learning WSI level representation by aggregating all patches embeddings via a WSI classifier with Recurrent Neural Network (RNN) [7] or in an attention [19] mechanism. The latter shows superior performance since such modeling includes less inductive bias compared to the former processing with fixed weights.

RNN [7] treats WSI as a sequence of patches, but its permutation variation and defects on long-term dependency do not match the nature of WSI well. By contrast, attention-based MIL (AB-MIL) [19] learns the weights of instance representations adaptively, and these methods can be flexibly inserted into the classification framework as a plug-and-play module to generalize the full-field map according to the task. To make patches' representation more discriminative, CLAM [30] introduces an auxiliary task in the MIL framework to distinguish its corresponding availability according to the size of instance attention while training the WSI classification. Trans-MIL [38] is designed to model the relation among patches via Self-Attention [41] and solves the softmax computational complexity in the long sequence by linear Self-Attention [42, 43]. To embed the multi-scale information of WSI, DS-MIL [25] simply concatenates patch features in three scales, and HIPT [9] builds hierarchical multi-scale features. To mitigate overfitting in complex attention model trained on the limited number of WSIs, DTFD-MIL [45] resamples instances from the original bag to generate various sub-bags for augmentation. However, all these works pay too much attention on WSI head to explore the backbone.
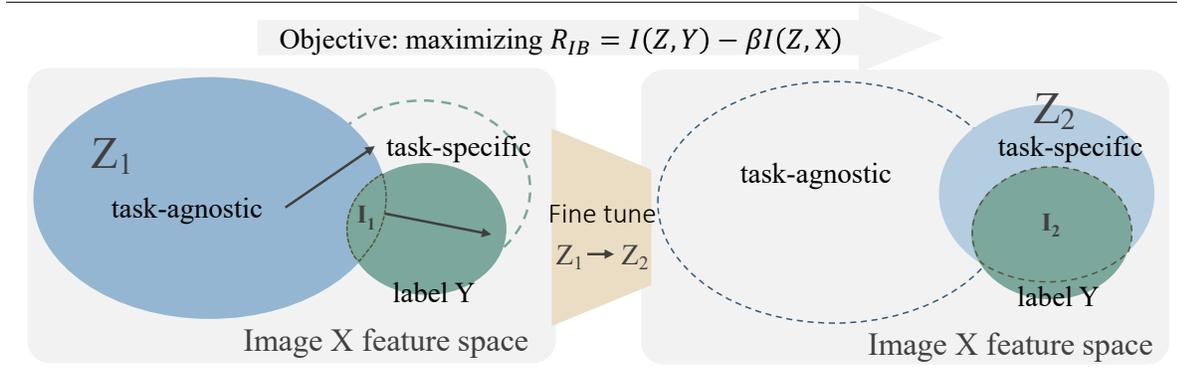
Figure 2. An illustration of the tuning scheme. To maximize R_{IB}, enlarging the first term helps latent representation vector Z become more predictive for label Y (More overlap between Y and Z2 compared to Z1, which denotes higher mutual information), while by reducing the second term to filter irrelevant features (smaller overlap between image X and Z2 compared to Z1).

2.2. Self-supervised Learning and Fine-tuning

Self-supervised Learning (SSL) has shown to be a promising paradigm both in computer vision [4,8,11,16,17] and natural language processing [12,33,44]. However, fine-tuning (FT) is quite necessary for the pretrained model of SSL to the downstream tasks, e.g. in [16,17], the vanilla FT method can improve about 15 percent of accuracy in IN-1K compared to vanilla Linear Probing [17]. More interesting and parameter-efficient FT methods [14,18,24,26,27,47,48] are proposed according to the similarity between pretraining and downstream in natural language processing and few-shot image generation. Unfortunately, the computational limitation on redundant instances prevents FT from being directly introduced into WSI analysis. Although SSL achieves optimistic performance in some works [9, 25] for pathology WSI analysis but only shows a small margin of improvement. There may be much potential to be discovered in WSI analysis if an appropriate FT method can be used to address the training cost from surplus information of WSI.

2.3. Information Bottleneck and Attributions

The Information Bottleneck (IB) conception is introduced in [40] as an information-theoretic framework for learning, which is currently used in deep learning both theoretically and practically. [39] proposes to unveil deep neural network in an information flow perspective by estimating Mutual Information (MI) between the outputs of two layers, then in followed-up works [15,35], they revisit the MI compression process of IB to propose complements and better measurements of MI in Deep Neural Network (DNN).

IB not only reveals how the traditional DNN training finishes information compression implicitly in the above works, but can also be employed explicitly as an objective to intervene in DNN's training: [2] derives a variational bound of the IB objective, claiming that the proxy objective provides an additional regularization term which improves the robustness to adversarial samples compared to the traditional training of DNN. [32] adopts the variational bound in [2] and replaces the distribution of latent features from Gaussian into Bernoulli. In this way, it generates masks to control the conciseness of the rationale extraction of sentences. Achille et al. [1] claim that IB solutions approximate minimal sufficient statistics (MSS) [23], and by penalizing the redundancy of representations, they heuristically argue that the model's sensitivity to nuisances is mitigated. Similar to some visual attribution methods like CAM [37, 49], LRP [3] and patch masking [4, 16, 28] in ViT, an IB-based attribution method is proposed in [36] by adding noise to intermediate feature maps, restricting the flow of information, then how much information image regions provide can be quantified. Intuitively, the IB objective reveals the limitation of pretraining for downstream tasks and the imperative of FT for task-specific representation in WSI analysis, as shown in Figure 2. Moreover, it motivates us to solve the dilemma of FT and computational limitation by its' MSS and visual attribution properties.

3. Method

3.1. Overview on MIL-based WSI Analysis

Given a WSI X, the goal is to make slide-level prediction Y by learning a classifier $f(X;\theta)$. Due to its extremely high resolution, X is patched into a huge bag of small instances $X = \{x_1, ..., x_N\}$, where N is the number of instance. The slide-level supervision $\hat{Y}$ is given by a Max-pooling operation of the latent label $\hat{y}_i$ for each instance x_i, which can be defined as:

$$\hat{Y} = \max\{\hat{y_1}, ..., \hat{y_N}\}. \tag{1}$$

Since all latent labels of instances $\hat{y}_i$ are unknown under the WSI-level supervision, conventional approaches con-

vert this problem into a MIL formulation in the following two steps: 1) Processing images into feature representations $Z = \{z_1, ..., z_N\}$ with a backbone h as $z_i = h(x_i; \theta_1)$ where h is a model of any architecture such as CNN or ViT with parameters θ_1. 2) Aggregating all patches' features within a slide and producing the slide-level prediction $Y = g(Z; \theta_2)$, where g is an attention-based pooling function followed by a linear classifier head as:

$$g(Z; \theta_2) = \sigma(\sum_{i=1}^{N} a_i z_i), \qquad (2)$$

where a_i is attention weights and $\sigma(\cdot)$ is a linear head. Limited by the computational cost, the parameters θ_1 and θ_2 in $f(X; \theta) = g\{h(X; \theta_1); \theta_2\}$ are learned separately by following steps: 1) Initializing θ_1 from the pretrained model, which refers to general features from public IN-1K, or learned by SSL on the related dataset to extract the domain-specific representations. 2) Freezing θ_1 and learning θ_2 under slide-level supervision.

3.2. Information Bottleneck for MIL Sparsity

Background of the Information Bottleneck

The Information Bottleneck (IB) can work as an information compression role to intervene in DNN's training [2]. The objective function of IB to be maximized is given in [40] as,

$$R_{IB} = I(Z, Y) - \beta I(Z, X), \qquad (3)$$

where $I(\cdot, \cdot)$ indicates the Mutual Information (MI) and β is a Lagrange multiplier controlling the trade-off between the information that the representation variable Z shares with the label Y and its shares with input X. Since the computation of MI is intractable during the training of the neural networks, to maximize IB objective can be transferred to minimize a variational bound of Eq.(3) derived in [2] follows:

$$J_{IB} = \frac{1}{N} \sum_{n=1}^{N} \mathbb{E}_{z \sim p_\theta(z|x_n)}[-\log q_\phi(y_n|z)] + \\ \beta KL[p_\theta(z|x_n), r(z)], \qquad (4)$$

where N denotes the number of samples, $q_\phi(y|z)$ is a parametric approximation to the likelihood $p(y|z)$, $r(z)$ is the prior probability of z to variational approximate the marginal $p(z)$, and $p_\theta(z|x)$ is the parametric posterior distribution over z.

Learn MIL Sparsity via Variational Bound of IB

To trade off the dilemma of computational limitation and task-specific representation learning via end-to-end back-propagation, we propose to utilize the IB module to filter most task-irrelevant instances for task-specific fine-tuning.

The above filtering process can be implemented by optimizing the second term of in Eq.(3) which controls the

compression. There are two ways that compress X to Z by decreasing the KL divergence between $p(z|x)$ and $r(z)$ in Eq.(4) variational method: reducing the dimension of representation Z compared to X in [2], or converting input X into a sparse one in [32].

For the setting of our long instance sequenced MIL, we reduce $I(X, Z)$ into a degree so that the gradients can be back-propagated to the backbone encoder, which needs us to convert a WSI of bag size over 10k into 1k for the sake of sparsity. Considering MIL for tumor v.s. normal binary classification without loss of generality and the latent label y_i of each instance x_i in Eq.(1), we argue that it is sufficient enough to make the WSI level prediction if one tumor area is detected. With the above understanding, we propose to learn compressed components similar to [32] by defining a IB module as:

$$z = m \odot x, \qquad (5)$$

where m is a Bernoulli(π) distributed binary mask and in this way $KL[p_\theta(z|x), r(z)]$ in Eq.(4) can be decomposed as,

$$KL[p_\theta(m_i|x), r(m_i)] + \pi H(X), \qquad (6)$$

where $H(X)$ is the entropy of X, which can be omitted during the minimization due to its constant value. Please check Supplementary for above proof.

The Bernoulli(π) distribution for m fits the definition of MIL empirically: we can treat m as a latent weak prediction $\hat{y}$ describing whether the patch contains tumor or not, denoting $P_{set} = \{p(m_1|x_1), ..., p(m_N|x_N)\}$, then during inference Eq.(1) can be derived as:

$$\hat{Y} = \max\{P_{set}\} = \max\{P_{subset}\}, \qquad (7)$$

where $P_{subset} \in P_{set}$, generated by select top-K elements in P_{set}. The patch classifier trained with only slide-level supervision shows low accuracy [25], so we only use it to generate mask for sparse sub-bag and still utilize attention-based MIL on the sub-bag for decision making.

3.3. Loss Function and its Implementation

Derived from Eq.(4) and Eq.(6), our Variational IB module should be optimized with the following loss function,

$$loss = \frac{1}{N} \sum_{n=1}^{N} \mathbb{E}_{z \sim p_\theta(z|x_n)}[-\log q_\phi(y_n|z)] + \\ \beta KL[p_\theta(m|x_n), r(m)], \qquad (8)$$

where the first term is the task loss to learn task-specific features, and it can be treated as the cross entropy same to prior works [2, 32] in a sampling perspective. The second term is the information loss to filter out task-irrelevant instances by minimizing the Kullback-Leibler divergence between the distribution of mask and the prior Bernoulli.
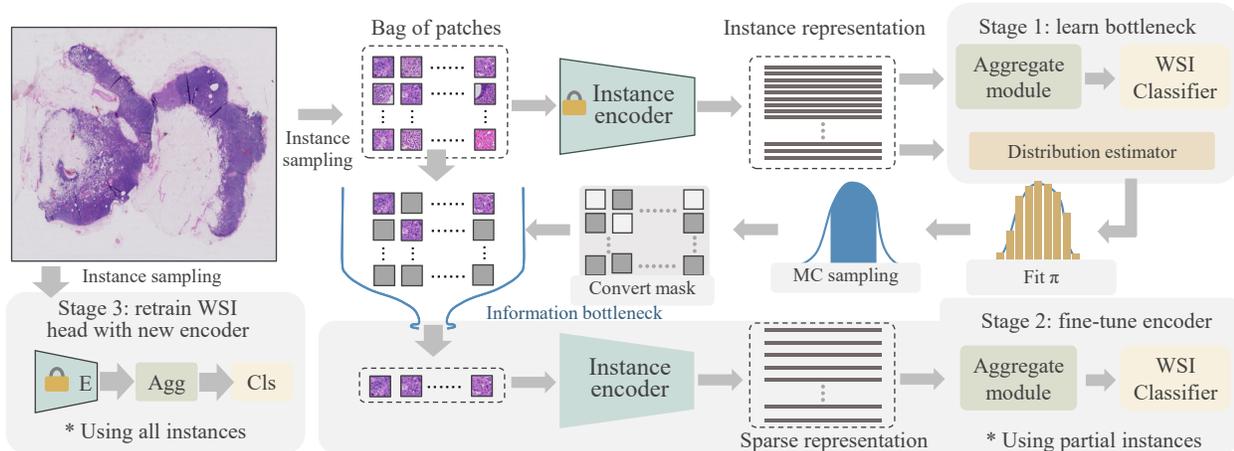
Figure 3. Workflow of WSI-MIL task-specific fine-tuning. 1) Initialize the backbone with pretrained parameters and set frozen, then learn the IB module to generate instance masks. 2) fix the mask to distill a sparse bag, then fine-tune the WSI head and patch the backbone end-2-end. 3) utilize all fine-tuned instance features within a bag and train the WSI-MIL classifier head.

The $p(m_n|x_n)$ is first generated by a linear layer from the representation of x_n and followed by a sigmoid activation. If the mask m is directly converted by thresholding $p(m_n|x_n)$, the gradients from $z = m \odot x$ can not be back-propagated since such an operation is not differentiable. So, during training, m is generated via Monte Carlo sampling, and $p(m_n|x_n)$ can be learned with the reparameterization trick [21, 22] for gradient's estimation.

3.4. Task-specific Fine-tuning

To realize task-specific fine-tuning and better performance with only slide-level labels, the workflow of our method is depicted in Figure 3, which includes 3 stages: 1) By minimizing the loss function in section 3.3 with a frozen backbone, a sparse instance set within a bag is obtained. 2) Assisted by the small portion of instances in a bag generated from stage 1, it is possible to fine-tune the backbone for better task-specific features and here we directly backpropagate the loss gradients from slide-level supervision into top-K patch instances. 3) Since the appropriateness of the sparse instance set heavily relies on the performance of the pretrained backbone, the first training IB module is insufficient and may result in small or even zero top-K recall. Moreover, the sparse instance set may lose the contextual dependency of WSI modeling. Thus in this stage, We utilize all fine-tuned instance features within a bag to train a traditional attention-based WSI-MIL classifier.

4. Experiments

In this section, we present the performance of the proposed method incorporated with the latest WSI-MIL frameworks, where the IN-1K and SSL features are used for com-

parisons. Ablation experiments are performed to further study the proposed method and for paper length, more experimental results are presented in the Supplementary.

Datasets and Tasks.

We use five datasets to evaluate our method. The slide-level classification performance of IN-1K, SSL, and our method is evaluated on three datasets including both histopathology and cytopathology images: two public histopathological WSI datasets, Camelyon-16 [5] for tumor / normal binary classification, The Cancer Genome Atlas Breast Cancer (TCGA-BRCA) [34] for tumors subtyping. One internal cytopathology WSI dataset is introduced to validate the universality of our method on both histo- and cyto-pathology, which is Liquid-based Preparation cytology for Cervical Cancer's early screening (LBP-CECA).

Despite the original evaluation on the closed dataset, we also evaluate the generalization of our method on Camelyon-16-C (generated with random synthetic domain shift from Camelyon-16), and Camelyon-17 [29] from five different centers, which occurs frequently in practical pathological diagnosis and has been hindering the application of automatical WSI analysis to the real world. Details of five datasets can be found in the supplementary.

For pre-processing, we follow the operations in CLAM-SB [30] which mainly includes HSV, Blur, Thresholding, and Contours methods to localize the tissue regions in each WSI. Then non-overlapping patches with size 256×256 on the 20X magnification are extracted from the tissue regions.

Pretraining and Fine-tuning.

Our work mainly focuses on the method of fine-tuning. Because good pretrained initialization results in better fine-tuning performance, we employ mainstream pretraining methods: 1) ImageNet-1k (IN-1K) data pretraining. 2) SSL

Method	Camelyon-16		TCGA-BRCA		LBP-CECA	
	F1	AUC	F1	AUC	F1	AUC
Full Supervision	0.967±0.005	0.992±0.003	-	-	0.741±0.006	0.942±0.002
RNN-MIL [7]	0.834±0.017	0.861±0.021	0.776±0.035	0.871±0.033	-	-
AB-MIL [19]	0.828±0.013	0.851±0.025	0.771±0.040	0.869±0.037	0.525±0.017	0.845±0.002
DS-MIL [25]	0.857±0.023	0.892±0.012	0.775±0.044	0.875±0.041	-	-
CLAM-SB [30]	0.839±0.018	0.875±0.028	0.797±0.046	0.879±0.019	0.587±0.014	0.860±0.005
TransMIL [38]	0.846±0.013	0.883±0.009	0.806±0.046	0.889±0.036	0.533±0.006	0.850±0.007
DTFD-MIL [45]	0.882±0.008	0.932±0.016	0.816±0.045	0.895±0.042	0.569±0.026	0.847±0.003
FT+ CLAM-SB	0.911±0.017	0.956±0.013	0.845±0.032	0.935±0.027	0.718±0.010	0.907±0.005
FT+ TransMIL	**0.923±0.012**	**0.967±0.003**	0.848±0.044	0.945±0.020	0.720±0.024	0.918±0.004
FT+ DTFD-MIL	0.921±0.007	0.962±0.006	**0.849±0.027**	**0.951±0.016**	0.723±0.008	0.922±0.005
Mean-pooling	0.629±0.029	0.591±0.012	0.818±0.022	0.910±0.032	0.350±0.017	0.735±0.006
Max-pooling	0.805±0.012	0.824±0.016	0.644±0.179	0.826±0.096	0.636±0.064	0.893±0.019
KNN (Mean)	0.468±0.000	0.506±0.000	0.633±0.066	0.749±0.055	0.393±0.000	0.650±0.000
KNN (Max)	0.559±0.000	0.535±0.000	0.524±0.032	0.639±0.063	0.477±0.000	0.743±0.000
FT+ Mean-pooling	0.842±0.006	0.831±0.007	**0.866±0.035**	**0.952±0.018**	0.685±0.014	0.900±0.002
FT+ Max-pooling	**0.927±0.011**	**0.969±0.004**	0.852±0.043	0.948±0.019	**0.695±0.013**	**0.912±0.004**
FT+ KNN (Mean)	0.505±0.000	0.526±0.000	0.784±0.044	0.907±0.034	0.529±0.000	0.737±0.000
FT+ KNN (Max)	0.905±0.000	0.916±0.000	0.802±0.063	0.882±0.036	0.676±0.000	0.875±0.000

Table 1. **Slide-Level Classification** by using the IN-1K pre-trained backbone or the proposed fine-tuned (FT) in three datasets. **Top Rows.** Different MIL architectures are compared to select the top 3 SOTA methods to validate the transfer learning performance using the IN-1K pre-trained backbone or the FT. **Bottom Rows.** The competition of various traditional aggregation and feature evaluation methods by using pre-trained IN-1K or the FT.

pretraining with SimCLR [11], MoCo [17] and DINO [8]. 3) for Camelyon-16 and LBP-CECA, since there are comprehensive tumor area annotations, we use the annotation to pretrain the patch backbone, which acts as the upper-bound of our method.

Different from prior works backbone frozen after pretraining, we perform backbone fine-tuning with the method proposed in 3.4. In this stage, we fine-tune the backbone and WSI model end-to-end with 25 epochs using the AdamW optimizer, batch size of 1 for WSI with bag size of 512, a learning rate of 1e-5 for backbone and 1e-3 for WSI head. For the BN layers in ResNet, we turn it to eval mode to fix statistics during fine-tuning since we find that the distribution of the top-K instances in a bag is limited to estimate statistics because of the similarity among instances. Shallow layers of backbone are frozen since they only focus on morphological features while deep layer focus on semantics.

4.1. Slide-level Classification

Evaluation Metrics. For all the experiments, macro-AUC and macro-F1 scores are reported since all the 3 datasets are class imbalanced. For Camelyon-16, the official training set is randomly split into training and validation sets with a ratio of 9:1. The experiment is conducted 5 times and the results of the official test set are reported. For TCGA-BRCA, we perform the 10-folder cross-validation with the same running setting adopted in HIPT [9]. Besides the dataset LBP-CECA is randomly split with a ratio of 6:1:3 for training, validation, and testing. The experiment is conducted 5 times. The mean and standard variance values of performance metrics are reported for multi-runs or cross-validation runs.

Comparison with baselines. Classification results are summarized in Table 1. We first show full patch supervision results as an upper bound, then we directly evaluate several classic WSI-MIL methods, including RNN-MIL [7], AB-MIL [19], DS-MIL [25], CLAM-SB [30], TransMIL [38], DTFD-MIL [45]. All the WSI-MIL baselines across the three classification tasks suffer from relatively low performance due to the inappropriate backbone features from the pretrained ResNet-50 in IN-1K. Then we apply FT to the backbone to obtain features for WSI classification by three WSI architectures including CLAM-SB (pure global attention, no intersection among instances), TransMIL (self-attention for long sequence), and DTFD-MIL (multi-tier attention paths followed with aggregation, good for patch imbalance).

The results show clearly consistent improvements equipped with FT features under AUC metric. CLAM-SB

achieves a performance relative increase of 9.26%, 6.37%, 5.47% on Camelyon-16, TCGA-BRCA, and LBP-CECA respectively with vanilla FT based on IN-1K pretraining. In the same conditions, TransMIL and DTFD-MIL achieve new SOTAs on all three datasets and can obtain better relative improvement, especially in LBP-CECA with an 8.00% and 8.85% growth compared to CLAM-SB. The results show similar improvements under the F1 metric. For the bottom two rows of Table 1, we compare simple Mean/Max-pooling training and KNN evaluations.

The competitive results demonstrate that our proposed FT method in the backbone of the MIL framework can boost the performance of WSI classification, even for the most simple feature-level Mean/Max-pooling. In addition, task-specific features help us better unearth the properties of different WSI-MIL architectures. Apparently, the more complex architecture, TransMIL, and DTFD-MIL, can not perform desirable performance with the frozen pretrained parameters in LBP-CECE. In contrast, its capabilities can be further enhanced by fine-tuning the backbone for the defined target. Moreover, the effects of different feature aggregation methods vary depending on the task. The better performance is shown by using Max-pooling aggregation in Camelyon-16 containing smaller tumor areas, while Mean-pooling is better in TCGA-BRCA.

4.2. Combination of SSL and Fine-tuning

In this section, we further improve the slide-level classification performance by combining SSL with FT.
Evaluation Metrics. TCGA-BRCA is used for the evaluation of its task complexity in tumor subtyping with the same setting to 4.1.
Comparison with baselines. Since there is no apparent single semantic object in small histopathology patches, we mainly compare SSLs by Contrastive Learning or augmentation like MoCo [17] and DINO [8], and SimCLR [11]. We show the results of SSLs performed in previous works [9, 10] for fair comparison. All results are performed with the same WSI architecture of CLAM-SB [30].

Experimental results are summarized in Table 2. Compared to vanilla IN-1K, MoCo and DINO (combined with FT) achieve consistent growth of 7.85% and 7.39 %, respectively. Besides, they have a slight increase of 1.39% and 0.96 % in comparison to IN-1K with FT. By SSL, the intrinsic task-agnostic features are learned from all patches in WSIs, and after applying our proposed FT for all modules, task-specific features can be distilled from the label and partial data. In such an SSL with an FT paradigm, the data and the label are explored comprehensively to produce a state-of-the-art WSI analysis.

4.3. Generalization on Domain Shift

In this section, we evaluate the generalization of slide-level classification models on domain shift [46] [20], which is crucial for real-world clinical applications due to the diversity in staining, preparation, and imaging devices for pathological image processing among hospitals.
Evaluation Metrics. The Camelyon-16-C dataset is generated from Camelyon-16 test set by synthetic domain shift, which adopts a random combination of Brightness, JPEG, and Hue proposed in [46]. For Camelyon-17, since it shares similarities to Camelyon-16 but is collected from five different medical centers, we randomly collect 30 samples from each center to evaluate the robustness of natural domain shift. We directly evaluate models on Camelyon-16-C and 150 extra data from Camelyon-17 with five times of running same as the previous setting in section 4.1. The standard variance values of results is omitted due to the constraint on

Method	F1	AUC
IN-1K [§]	-	$0.884_{\pm 0.059}$
IN-1K	$0.797_{\pm 0.046}$	$0.879_{\pm 0.019}$
/w FT	$0.845_{\pm 0.032}$	$0.935_{\pm 0.027}$
SimCLR [11][§]	-	$0.879_{\pm 0.069}$
MoCo [17]	$0.804_{\pm 0.042}$	$0.904_{\pm 0.030}$
/w FT	$0.851_{\pm 0.029}$	$0.948_{\pm 0.026}$
DINO [8] [§]	-	$0.886_{\pm 0.059}$
DINO	$0.801_{\pm 0.045}$	$0.891_{\pm 0.043}$
/w FT	$0.848_{\pm 0.027}$	$0.944_{\pm 0.036}$

Table 2. **Combination of SSL and Fine-tuning.** We compare SSLs with IN-1K and their further improvement via fine-tuning (FT) on TCGA-BRCA. The symbol [§] indicates the result released in previous publication [9, 10].

Method	Camelyon-16-C		Camelyon-17	
	F1	AUC	F1	AUC
Max-pooling	0.689	0.742	0.578	0.670
/w FT	0.816	0.892	0.687	0.720
CLAM-SB [30]	0.742	0.836	0.624	0.702
/w FT	0.823	0.862	0.676	0.725
TransMIL [38]	0.748	0.842	0.657	0.706
/w FT	0.795	0.857	0.684	0.717
DTFD-MIL [45]	0.775	0.799	0.576	0.676
/w FT	0.804	0.838	0.689	0.717

Table 3. **Generalization on Domain Shift.** The generalization ability of all methods is compared between fine-tuning(FT) and IN-1K features on two datasets with domain shift. Camelyon-16-C and Camelyon-17 are synthetic and real corruptions respectively.

the paper length and a more clear comprehensive comparison can be found in Supplementary.

Comparison with baselines. The evaluation of generalization results is summarized in Table 3. For WSI model architecture, despite CLAM-SB [30], TransMIL [38], DTFD-MIL [45], we also compare Max-pooling on feature level since it shows promising results on Camelyon-16 in 4.1.

The results demonstrate consistent improvements after equipping the WSI head with fine-tuned features under the F1 metric: In general, all models can resist domain shifts to a certain extent by utilizing the proposed fine-tuning. Compared with freezing IN-1K parameters, Max-pooling obtains a noticeable performance increase of 18.43% and 18.86% on Camelyon-16-C and Camelyon-17, respectively. CLAM-SBachieves an increase of 10.91%, 8.33% respectively. TransMIL achieves an increase of 6.28%, 4.11% respectively. DTFD-MIL achieves an increase of 2.90%, 11.30% respectively. In contrast, the generalization of DTFD-MIL can still be further improved by our proposed FT optimization, even though it is already robust to different domains.

Interestingly, all improvements in performance under the AUC metric are much less than F1. The three Attention-based pooling methods respectively achieve the average improvement of 6.70% and 7.91% on Camelyon-16-C and Camelyon-17 under the F1 metric, as compared to 2.93% and 2.98% under AUC, which indicates that the classification ability of WSI-MIL model (measured by F1) may be much weaker than ranking (measured by AUC). However, doctors may much more care about F1 instead of AUC in clinical diagnosis from sensitivity and specificity perspectives. Most importantly, the classification threshold usually keeps fixed after deploying the model in the real world, which reflects the contribution of our method is more meaningful in practical applications.

4.4. Further Ablation Experiments

Additional ablations are included in the Supplementary Materials, with the main focus on the effects of learning rate on the backbone, Number selection of Top-K, and Value selection of Lagrange multiplier, results of the 3 stages.

4.5. Interpretability and Visualization

Here, we further show the interpretability improvements with FT features. As shown in Figure 4, attention scores from CLAM-SB [30] were visualized as a heatmap to determine the ROI and interpret the important morphology used for diagnosis. Obviously, the model focuses more on tumors with FT features.

5. Conclusion

In this work, we present a fine-tuning method for WSI classification under the weak supervision of slide-level la-
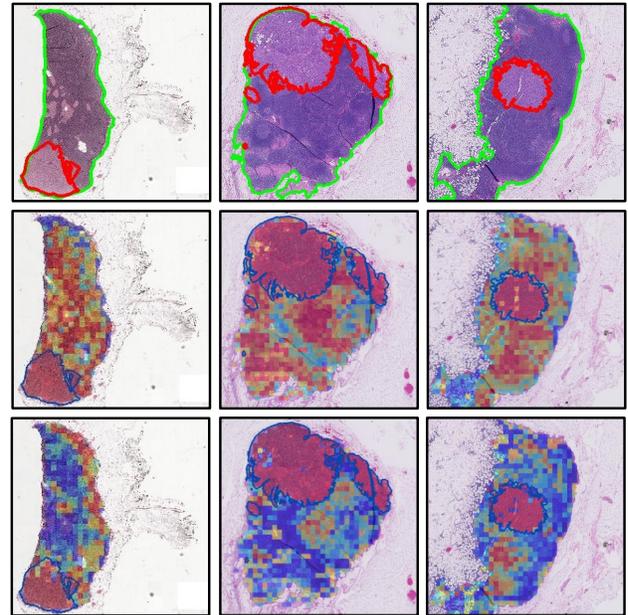

Figure 4. Heatmap comparison between pretraining and FT. The first row shows the full annotations, where red contours denote the tumor area and green contours filter the nonsense background. The second and third row shows the WSI attention map on pretraining and FT features respectively.

bels. Initially, an effective IB module is introduced to mitigate the training cost of Gigapixel WSI, which distills the oversized bag into a sparse one. Then the backbone of the instance is able to be trained end-to-end in the MIL framework by learning and making classifications on the distilled bag. Thus the WSI classification performance is improved through the retention of mainly task-specific information. In addition, SSL can be combined with the proposed framework for further improvement. Compared with fully supervised learning, our methods can achieve competitive accuracy by utilizing extremely weak WSI labels. Furthermore, our training scheme can introduce versatile training-time augmentations for better generalization on datasets with domain shift, which is an inevitable challenge for previous work. The experimental results reflect the advances of our method in both accuracy and generalization. Overall, Our proposed approach shows strong potential for MIL application in real-world pathology diagnosis, with better performance, faster convergence and annotation efficiency.

Acknowledgement

References

[1] Alessandro Achille and Stefano Soatto. Information dropout: Learning optimal representations through noisy computation. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2897–2905, 2018. 2, 3

[2] Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2017. 2, 3, 4

[3] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015. 3

[4] Hangbo Bao, Li Dong, and Furu Wei. Beit: BERT pre-training of image transformers. *CoRR*, abs/2106.08254, 2021. 3

[5] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, Meyke Hermsen, 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. 5

[6] Wouter Bulten, Kimmo Kartasalo, Po-Hsuan Cameron Chen, Peter Ström, Hans Pinckaers, Kunal Nagpal, Yuannan Cai, David F Steiner, Hester van Boven, Robert Vink, et al. Artificial intelligence for diagnosis and gleason grading of prostate cancer: the panda challenge. *Nature medicine*, 28(1):154–163, 2022. 1

[7] Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, 25(8):1301–1309, 2019. 2, 6

[8] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *CoRR*, abs/2104.14294, 2021. 2, 3, 6, 7

[9] Richard J. Chen and et al. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *CVPR*, pages 16144–16155, June 2022. 2, 3, 6, 7

[10] Richard J Chen and Rahul G Krishnan. Self-supervised vision transformers learn visual concepts in histopathology. *arXiv preprint arXiv:2203.00585*, 2022. 7

[11] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. 2, 3, 6, 7

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018. cite arxiv:1810.04805Comment: 13 pages. 3

[13] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes van Diest, Bram van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen A. W. M. van der Laak, , and the CAME-

[14] LYON16 Consortium. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. *JAMA*, 318(22):2199–2210, 12 2017. 1

[14] Luciano Floridi and Massimo Chiriatti. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30(4):681–694, 2020. 3

[15] Ziv Goldfeld, Ewout van den Berg, Kristjan Greenewald, Igor Melnyk, Nam Nguyen, Brian Kingsbury, and Yury Polyanskiy. Estimating information flow in deep neural networks. *arXiv preprint arXiv:1810.05728*, 2018. 3

[16] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. *CoRR*, abs/2111.06377, 2021. 2, 3

[17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735, 2020. 2, 3, 6, 7

[18] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 3

[19] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2127–2136. PMLR, 10–15 Jul 2018. 1, 2, 6

[20] Syed Ashar Javed, Dinkar Juyal, Zahil Shanis, Shreya Chakraborty, Harsha Pokkalla, and Aaditya Prakash. Rethinking machine learning model evaluation in pathology. *arXiv preprint arXiv:2204.05205*, 2022. 7

[21] Durk P Kingma, Tim Salimans, and Max Welling. Variational dropout and the local reparameterization trick. *Advances in neural information processing systems*, 28, 2015. 5

[22] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 5

[23] Erich Leo Lehmann and Henry Scheffé. Completeness, similar regions, and unbiased estimation-part i. In *Selected works of EL Lehmann*, pages 233–268. Springer, 2012. 2, 3

[24] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021. 3

[25] 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. 2, 3, 4, 6

[26] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021. 3

[27] Yijun Li, Richard Zhang, Jingwan Lu, and Eli Shechtman. Few-shot image generation with elastic weight consolidation. *arXiv preprint arXiv:2012.02780*, 2020. 3

[28] Youwei Liang, Chongjian Ge, Zhan Tong, Yibing Song, Jue Wang, and Pengtao Xie. Not all patches are what you need: Expediting vision transformers via token reorganizations. *arXiv preprint arXiv:2202.07800*, 2022. 3

[29] Geert Litjens, Peter Bandi, Babak Ehteshami Bejnordi, Oscar Geessink, Maschenka Balkenhol, Peter Bult, Altuna Halilovic, Meyke Hermsen, 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. 5

[30] 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. 2, 5, 6, 7, 8

[31] Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. *Advances in neural information processing systems*, 10, 1997. 1

[32] Bhargavi Paranjape, Mandar Joshi, John Thickstun, Hannaneh Hajishirzi, and Luke Zettlemoyer. An information bottleneck approach for controlling conciseness in rationale extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1938–1952, Online, Nov. 2020. Association for Computational Linguistics. 3, 4

[33] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. 3

[34] Nicholas A Petrick, Shazia Akbar, Kenny HH Cha, Sharon Nofech-Mozes, Berkman Sahiner, Marios A Gavrielides, Jayashree Kalpathy-Cramer, Karen Drukker, Anne LL Martel, et al. Spie-aapm-nci breastpathq challenge: an image analysis challenge for quantitative tumor cellularity assessment in breast cancer histology images following neoadjuvant treatment. *Journal of Medical Imaging*, 8(3):034501, 2021. 5

[35] Andrew M Saxe, Yamini Bansal, Joel Dapello, Madhu Advani, Artemy Kolchinsky, Brendan D Tracey, and David D Cox. On the information bottleneck theory of deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124020, 2019. 3

[36] Karl Schulz, Leon Sixt, Federico Tombari, and Tim Landgraf. Restricting the flow: Information bottlenecks for attribution. *arXiv preprint arXiv:2001.00396*, 2020. 3

[37] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. 3

[38] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, and yongbing zhang. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 2136–2147. Curran Associates, Inc., 2021. 2, 6, 7, 8

[39] Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017. 3

[40] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000. 3, 4

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 2

[42] Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020. 2

[43] Yunyang Xiong, Zhanpeng Zeng, Rudrasis Chakraborty, Mingxing Tan, Glenn Fung, Yin Li, and Vikas Singh. Nyströmformer: A nyström-based algorithm for approximating self-attention. 2021. 2

[44] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *NeurIPS*, 2019. 3

[45] Hongrun Zhang, Yanda Meng, Yitian Zhao, Yihong Qiao, Xiaoyun Yang, Sarah E. Coupland, and Yalin Zheng. Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. *ArXiv*, abs/2203.12081, 2022. 1, 2, 6, 7, 8

[46] Yunlong Zhang, Yuxuan Sun, Honglin Li, Sunyi Zheng, Chenglu Zhu, and Lin Yang. Benchmarking the robustness of deep neural networks to common corruptions in digital pathology. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 242–252. Springer, 2022. 7

[47] Yunqing Zhao, Keshigeyan Chandrasegaran, Milad Abdollahzadeh, and Ngai-Man Cheung. Few-shot image generation via adaptation-aware kernel modulation. *arXiv preprint arXiv:2210.16559*, 2022. 3

[48] Yunqing Zhao, Henghui Ding, Houjing Huang, and Ngai-Man Cheung. A closer look at few-shot image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9140–9150, June 2022. 3

[49] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, 2016. 3