LNPL-MIL: Learning from Noisy Pseudo Labels for Promoting Multiple Instance Learning in Whole Slide Image

Zhuchen Shao[1], Yifeng Wang[2], Yang Chen[1], Hao Bian[1], Shaohui Liu[3], Haoqian Wang[1*], Yongbing Zhang[2*]
[1]Tsinghua Shenzhen International Graduate School, Tsinghua University
[2]Harbin Institute of Technology (Shenzhen) [3]Harbin Institute of Technology
shaozc0412@gmail.com, ybzhang08@hit.edu.cn

Abstract

Gigapixel Whole Slide Images (WSIs) aided patient diagnosis and prognosis analysis are promising directions in computational pathology. However, limited by expensive and time-consuming annotation costs, WSIs usually only have weak annotations, including 1) WSI-level Annotations (WA) and 2) Limited Patch-level Annotations (LPA). Currently, Multiple Instance Learning (MIL) often exploits WA, while LPA usually assign pseudo-labels for unlabeled data. Intuitively, pseudo-labels can serve as a practical guide for MIL, but the unreliable prediction caused by LPA inevitably introduce noise. Furthermore, WA-supervised MIL training inevitably suffers from the semantical unalignment between instances and bag-level labels. To address these problems, we design a framework called Learning from Noisy Pseudo Labels for promoting Multiple Instance Learning (LNPL-MIL), which considers both types of weak annotation. Specifically, for the LPA-trained weak classifier, we design a Super-Patch-based LNPL (SP-LNPL) method to reduce false positives in the noisy pseudo-labels and then select more accurate Top-K key instances. In MIL, we propose a Transformer aware of instance Order and Distribution (TOD-MIL) that strengthens instances correlation and weakens semantical unalignment in the bag. We validate our LNPL-MIL on Tumor Diagnosis and Survival Prediction, achieving state-of-the-art performance with at least 2.7%/2.9% AUC and 2.6%/2.3% C-Index improvement with the patches labeled for two scale. Ablation study and visualization analysis further verify the effectiveness.

1. Introduction

In *Computational Pathology* (CPATH), limited by the high-resolution, wide-field of view property (about 50,000 $\times$ 50,000 pixels) of *Whole Slide Images* (WSIs) [46] and the biomedical backgrounds required for data annotations,

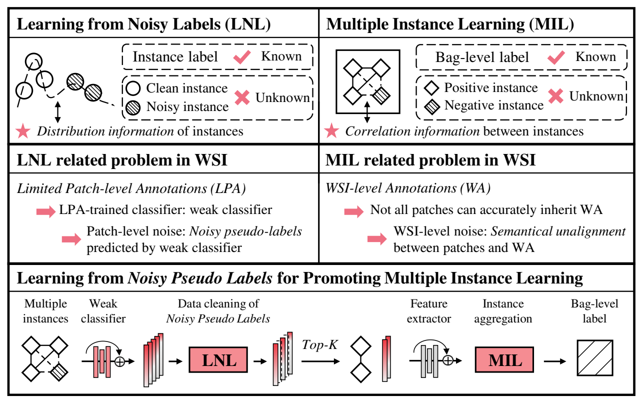


Figure 1. **Background and Motivation. Top Row.** Definition of LNL and MIL. **Second Row.** LNL and MIL related problems in WSI. **Bottom Row.** LNL assists MIL training with Top-K key instances selection and more accurate instance pseudo-labels.

WSIs usually only have two types of weak annotation [26] : **1)** *WSI-level Annotations* (WA); **2)** *Limited Patch-level Annotations* (LPA). Currently, *Multiple Instance Learning* (MIL) is often used to address vital WSI-level tasks in CPATH, *e.g.*, cancer subtype diagnosis, patient prognosis, therapeutic-response prediction, and biomarker prediction [8,13,33,34,39]. As shown in Fig. 1, applying weak annotations will inevitably introduce noise. Designing the framework to facilitate WSI-level tasks with both types of weak annotation is still a challenging problem in CPATH.

Intuitively, LPA can help WA-supervised MIL training explore the correlation information between instances and help select Top-K key instances. Unfortunately, the methods for utilizing LPA, such as *Fully Supervisd Learning* (FSL) [1, 58] and *Semi-Supervised Learning* (SSL) method [17, 18, 36], still cannot get satisfactory results. Specifically, only weak classifiers can be trained with LPA, and the pseudo-labels usually contain a lot of noise. Besides, when the labeled and unlabeled data come from different centers, differences in tissue preparation and scanners

*Corresponding authors.

further amplify the effect of patch-level noise [6, 25]. As a *Learning from Noisy Labels* (LNL) problem described in Fig. 1, an effective *Learning from Noisy Pseudo Labels* (LNPL) method is urgently needed in LPA utilization.

Furthermore, in WA-supervised MIL training, not all patches can accurately inherit WA. Here, we discuss two typical scenarios: **1)** Partial patches can independently inherit WA, *e.g.*, for the Tumor Diagnosis in Camelyon16 [5], only a tiny percentage of tumor patches can inherit WA. **2)** No patches can independently inherit the WA, *e.g.*, for the Survival Prediction in TCGA [32], only the joint representation of many patches, such as the tumor microenvironment, can inherit WA. In this paper, we describe this label ambiguity [9] as a *semantical unalignment* between patches and WSI-level label. To address this problem, on the one hand, some studies combine MIL assumptions [8, 27, 57], or prior medical knowledge [1, 24] to select Top-K key instances that can semantically align with the bag-level label. On the other hand, some studies employ feature-based end-to-end training, *i.e.*, bag-level labels directly guide feature aggregation so that the semantically unaligned feature will be paid less attention. Among them, various forms of attention such as bypass attention [11, 33, 34, 62], non-local attention [28], and self-attention [14, 16, 40, 41], are widely used. Besides, spatial information and long-distance dependencies in the bag [22, 29, 40] are also widely explored. Currently, most studies explore the ideal case that only WA participates in MIL training. A promising direction is how to combine LPA to promote MIL training, *i.e.*, strengthen instance correlation and weaken semantical unalignment.

Based on two common weak annotation forms: WA and LPA, we design a framework called *Learning from Noisy Pseudo Labels for promoting Multiple Instance Learning* (LNPL-MIL). The contributions are as follows:

1) We verify the superiority of the LNPL-MIL under two representative WSI-level prediction tasks: Tumor Diagnosis and Survival Prediction. Compared to a series of competing methods, our LNPL-MIL framework achieves *State-Of-The-Art* (SOTA) performance: at least 2.7%/2.9% AUC and 2.6%/2.3% C-Index improvement can be achieved with the patches labeled for two scale.

2) We design a *Super-Patch-based LNPL* (SP-LNPL) method to select more accurate Top-K key instances. SP-LNPL jointly leverages global feature distribution and the LPA-trained weak classifier with the FSL, which can efficiently reduce false positives. Compared with the FSL and even SOTA SSL methods, SP-LNPL achieves *higher metrics* in both patch-level and WSI-level tasks.

3) We propose a *Transformer aware of instance Order and Distribution* (TOD-MIL). By strengthening instances correlation and weakening semantical unalignment in the bag, we fully utilize WA and LPA to obtain *superior* performance in WSI-related downstream tasks.

2. Related Work

2.1. Learning from Noisy Labels in WSI.

As shown in Fig. 1, the LNL problem can be defined as the existence of a set of training instances, the labels of the instances are known, but it is unknown whether the instance labels are noisy. The LNL problem explores the distribution of the clean instances to remove noisy labels. The methods of LNL in natural images include robust model architecture design [55, 59], regularization [48, 54], loss function [20, 43], and screening of training samples [45, 50]. As gigapixel images, WSIs are always influenced by noisy labels. Wang *et al*. [51] discuss the noisy labels caused by coarse annotations in WSIs and propose a MIL-based denoising method, which can achieve better results than the deep KNN method [4]. For the inaccuracy and incomplete labeling problems, the superpixels method [2, 7] and SSL method [17, 18, 30] are applied to reduce the interference of noisy labels. Unlike the noisy annotation problem discussed above, this paper will discuss the *noisy pseudo-labels* caused by LPA-trained weak classifiers. Further, learning from noisy pseudo-labels for promoting MIL.

2.2. Multiple Instance Learning in WSI.

As shown in Fig. 1, the MIL problem can be defined as the existence of a set of training instances, the overall label of the instances (bag-level label) is known, but the labels of each instance are unknown. The MIL problem explores how to aggregate a bag of unlabeled instances to predict bag-level labels. Limited by the difficulty of annotation in WSI, MIL, as a weakly supervised learning method, has been widely used in WSI-related tasks. There are currently two mainstream studies, including image-level Top-K key instances selection [8, 15, 27, 57], feature-level instances aggregation [21, 28, 34, 40, 62] and some composite variants [37, 38, 56]. Currently, benefiting from the long-distance communication ability, many MIL models [10, 12, 16, 29, 40] adopt the Transformer to explore correlation information between instances within a bag. Most MIL methods in WSIs are designed when only WA exists. Some studies also discuss a more realistic situation: LPA are also accessible. Bian *et al*. [7] adopt the superpixels method and mixed supervision strategy to jointly use LPA (*e.g.*, Gleason pattern) and WA (*e.g.*, ISUP grade). Gao *et al*. [17] propose a semi-supervised multi-task learning framework to cooperate the weak annotation including LPA (*e.g.*, min-point annotation) and WA (*e.g.*, cancer subtyping). However, the current methods are only validated on cancer classification tasks with large tumor areas. More general and effective methods are still worth exploring in more complex scenarios, *e.g.*, small tumor area, LPA and WA from different centers, and regression problems such as survival prediction.

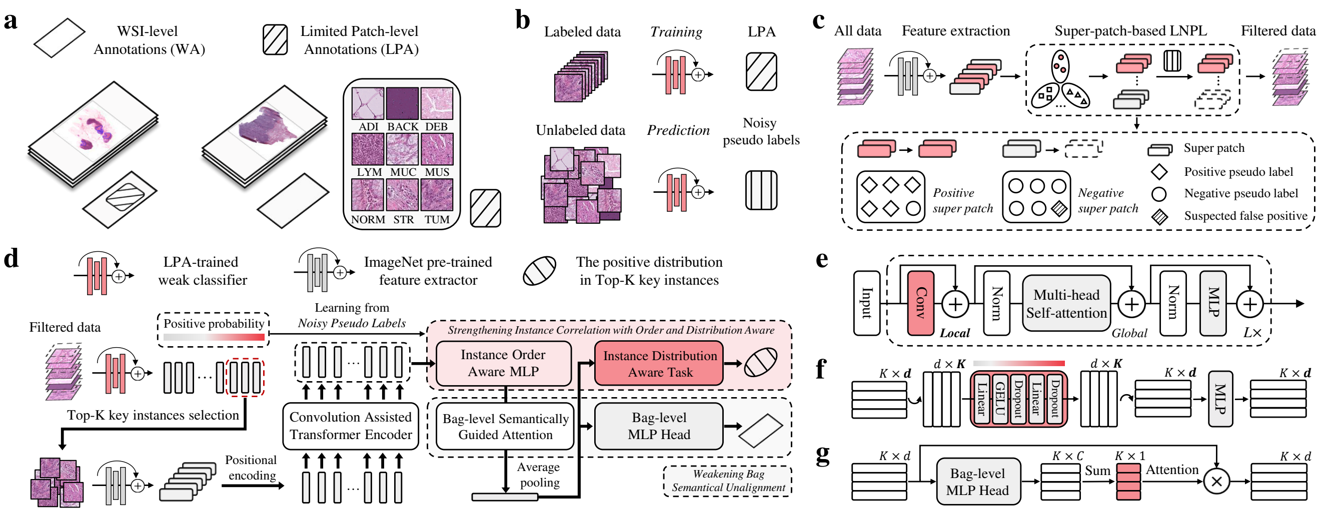


Figure 2. **Overview of LNPL-MIL Framework. a.** Two typical scenarios for WSI-related tasks. **b.** We first use LPA to train a weak classifier, then employ this classifier to assign pseudo-labels to unlabeled data. **c.** The *Super-Patch-based LNPL* (SP-LNPL) method looks for the ROI super patches in the feature space to filter the suspected false positives. **d.** In MIL, we propose a *Transformer aware of instance Order and Distribution* (TOD-MIL). It includes *Convolution Assisted Transformer Encoder* (C-Trans), *Instance Order Aware MLP* (IOA-MLP), *Bag-level Semantically Guided Attention* (BG-Attn), and *Instance Distribution Aware Task* (IDA-Task). **e.** The specific form of C-Trans. **f.** The specific form of IOA-MLP. **g.** The specific form of BG-Attn.

3. Method

3.1. Problem Formulation

In CPATH, two common forms of annotation are WA and LPA. MIL is often used to solve problems containing only WA. For a bag of instances $\mathcal{X} = \{(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_n, y_n)\}$, bag-level (WSI-level) label Y is known, but instance (patch-level) label y is unknown. Actually, a small number of patch-level annotations are sometimes accessible, which is the so-called LPA. So we further define y_l to represent labeled y and y_u to represent unlabeled y. It is worth noting that the annotations Y and y_l do not always come from the same dataset. As shown in Fig. 2a, we discuss two scenarios: **1)** The same dataset contains both Y and limited y_l; **2)** Dataset $\mathcal{A}$ contains Y, and Dataset $\mathcal{B}$ contains limited y_l.

3.2. Top-K Key Instances Selection with LNPL

As a high-resolution image, a WSI contains thousands of patches. Since not all patches can accurately inherit WSI-level labels, Top-K key instances selection is an efficient way to reduce unrelated patches and alleviate high computational costs. Intuitively, LPA can give guidance to select key instances. However, LPA-trained classifiers usually only have unreliable prediction, which introduces noise inevitably. In this section, we first give formal definitions for weak classifier and pseudo labels, and then describe the LNPL method, which help reduce noise in pseudo-labels and select more accurate Top-K key instances.

Weak Classifier Training and Pseudo-labels Assigning. As shown in Fig. 2b, we first use LPA to train a weak classifier. Then, the LPA-trained weak classifier is employed to assign pseudo-labels for the remaining unlabeled patches. Specifically, we choose a small model, ResNet18 [19]. To obtain a weak classifier with LPA, we simply apply the FSL training as the baseline. For the pseudo-labels assigning, it can be defined as follows:

$$\boldsymbol{y}_p = \mathrm{F}_{\text{weak}}\left(\boldsymbol{x}\right), \, \hat{y}_p = \arg\max \boldsymbol{y}_p, \tag{1}$$

where $\boldsymbol{y}_p$ denotes the positive probability, $\hat{y}_p$ denotes the assigned pseudo-label, $\boldsymbol{x}$ denotes the input patch, and F_{weak} denotes the weak classifier trained with LPA.

Super-patch-based LNPL Method. Constrained by LPA, the weak classifiers may incorrectly assign high/low positive probabilities to some negative/positive instances inevitably. In contrast, the KNN search is not affected by LPA. It can finish classification with the help of all data and find more similar patches in high dimensional space. Therefore, we can regard KNN search as a weak classifier that learns from the coarse-grained global distribution. Similarly, the LPA-trained weak classifier can be regarded as learning from fine-grained local distribution. Intuitively, combining two types of predictions leads to more accurate pseudo-labels. Since false positives are more prevalent in Top-K instances selection and are more harmful to WSI-level tasks, we focus on reducing false positives in the LNPL method design.

Currently, superpixels based clustering method [2, 7] is adopted to reduce the influence of noisy annotations. However, the superpixels method has the following problems. **1)** The clustering results based on image texture are relatively rough and the clustering range is limited within the local space; **2)** The sizes of different superpixels are inconsistent, and it isn't easy to quantitatively and fairly measure which superpixel has fewer false positives.

To address these problems, as shown in Fig. 2c, we design a *Super-Patch-based LNPL* (SP-LNPL) method. For Problem **1)**, since using the LPA-trained weak classifier to extract features will introduce the bias of limited annotations inevitably, we first employ the ImageNet pre-trained ResNet18 to embed all the patches into task-agnostic [47] features $\mathcal{H}$. Then, we employ the $\mathcal{H}$ to perform global KNN search [35] to classify similar patches into the same super patch. For Problem **2)**, we first divide the $\mathcal{H}$ into a series of same-sized super patches, then combining pseudo-labels to quantitatively and fairly compare false positives across different super patches. Besides, since the proportion of positive patches in different WSIs is distinct, we adopt the adaptive threshold method to reassign ROI pseudo-labels $\hat{y}_{roi}$ for patches in each WSI. Then, the positive ratio of $\hat{y}_{roi}$ is used to measure whether the super patch is an ROI super patch. Finally, we select the Top-K key instances with the largest positive probability from all remaining ROI super patches. The implementation is shown in Algorithm 1.

Algorithm 1 Super-patch-based LNPL Method

Input: The bag $\mathcal{X}$ with a series of instances $\{(\boldsymbol{x}_1, \boldsymbol{y}_{p,1}), \ldots, (\boldsymbol{x}_n, \boldsymbol{y}_{p,n})\}$. Each super patch size is w. The ratio threshold of ROI positive patches in each super patch is t_{ROI}.
Output: Top-K key instances $\hat{x}_1, \ldots, \hat{x}_K$.
%1. Feature extraction, applying the ImageNet pre-trained model.
for $i \in [1, n]$ **do** $\boldsymbol{h}_i \leftarrow \mathrm{F}_{\text{ImageNet}}(\boldsymbol{x}_i)$;
%2. Feature pre-processing, padding for feature sequence $\mathcal{H}$.
$\mathcal{H}_a,\ \mathcal{X}_a \leftarrow \text{Padding}\left((\mathcal{H},\ \mathcal{X})\right)$
%3. ROI pseudo label pre-processing.
$y_{\text{Mid}} \leftarrow \boldsymbol{y}_p\left[idx_{\text{Mid}}\right]$ ◁ Find the median positive probability
for $i \in [1, n]$ **do** $\hat{y}_{roi,i} \leftarrow 1$ **if** $\boldsymbol{y}_{p,i} > y_{\text{Mid}}$ **else** 0;
%4. KNN search, looking for ROI super patches.
Initialize $\hat{\mathcal{X}}$ as $\emptyset$
for $i \in [0 : N : w]$ **do**
  %4.1. Select the w features closest to $\boldsymbol{h}_1$ in $\mathcal{H}_a$.
  $idx \leftarrow \text{Hnsw.query}\left(\boldsymbol{h}_1, \text{topn} = w\right)$
  $\boldsymbol{x}_{idx},\ \boldsymbol{y}_{p,idx},\ \hat{y}_{roi,idx} \leftarrow \mathcal{X}_a\left(idx\right)$
  %4.2. Count the ratio of positive ROI pseudo-labels.
  $ratio \leftarrow \text{Count}\left(\hat{y}_{roi,idx} == 1\right)$
  %4.3. Determine whether the cluster is an ROI super patch.
  if $ratio > t_{ROI}$ **then** $\hat{\mathcal{X}} \leftarrow \hat{\mathcal{X}} + \mathcal{X}_a\left(idx\right)$ ◁ Update $\hat{\mathcal{X}}$;
  $\mathcal{X}_a \leftarrow \mathcal{X}_a - \mathcal{X}_a\left(idx\right)$ ◁ Delete selected idx from $\mathcal{X}_a$
end
%5. Select Top-K key instances from the filtered bag $\hat{\mathcal{X}}$.
$\hat{x}_1, \ldots, \hat{x}_K \leftarrow \text{Max}\left(\hat{\boldsymbol{y}}_{p,1}, \ldots, \hat{\boldsymbol{y}}_{p,n'}\right)$
return $\hat{x}_1, \ldots, \hat{x}_K$

3.3. Transformer Aware of Instance Order and Distribution in MIL

As we have described in the Sec. 3.2, the SP-LNPL aided Top-K key instances selection has the following advantages. **1)** The instances connection becomes more closely after the Top-K key instances selection, reflecting instance positive probability order; **2)** The pseudo-labels become more accurate after the SP-LNPL method, reflecting the instance positive distribution. Therefore, in the MIL training, we design a *Transformer aware of instance Order and Distribution* (TOD-MIL). Specifically, it mainly consists of two parts, *i.e*., strengthening instance correlation and weakening bag semantical unalignment.

3.3.1 Strengthening Instance Correlation with Instance Order and Distribution Aware

Convolution assisted Transformer Encoder. Sufficient interactions between instances are the basis for instance order and distribution exploration. To effectively facilitate local and global connections during feature aggregation, we introduce the *1D Convolution to the Transformer encoder* (C-Trans). Given the Top-K key instance features $\mathcal{H}_t^0 \in \mathbb{R}^{K\times d}$, the procedure can be defined as follows:

$$\begin{aligned}
\mathcal{H}_t^\ell &= \mathrm{Conv}(\mathcal{H}_t^{\ell-1}) + \mathcal{H}_t^{\ell-1},\ \ell = 1 \ldots L \\
\mathcal{H}_t^\ell &= \mathrm{MSA}(\mathrm{LN}(\mathcal{H}_t^\ell)) + \mathcal{H}_t^\ell,\ \ell = 1 \ldots L \\
\mathcal{H}_t^\ell &= \mathrm{MLP}(\mathrm{LN}(\mathcal{H}_t^\ell)) + \mathcal{H}_t^\ell,\ \ell = 1 \ldots L
\end{aligned} \tag{2}$$

where L is the number of layers, Conv denotes 1D Convolution, MSA denotes Multi-head Self-attention, MLP denotes Multilayer Perceptron, and LN denotes Layer Norm.

Instance Order Aware MLP. The C-Trans strengthens local-global connections to provide better feature representation, and Top-K key instances selection assisted with SP-LNPL provides more accurate order information. This implicit order connection among selected instances can effectively guide the TOD-MIL to learn the instance interactions. Specifically, inspired by the position-aware module proposed in [49], we design an *Instance Order Aware MLP* (IOA-MLP). Given the output of Transformer encoder $\mathcal{H}_t^L \in \mathbb{R}^{K\times d}$, the procedure can be defined as follows:

$$\begin{aligned}
\mathcal{H}_d &= \mathrm{MLP}({\mathcal{H}_t^L}^{\mathrm{T}}) + {\mathcal{H}_t^L}^{\mathrm{T}}, \\
\mathcal{H}_D &= \mathrm{MLP}({\mathcal{H}_d}^{\mathrm{T}}) + {\mathcal{H}_d}^{\mathrm{T}},
\end{aligned} \tag{3}$$

where $\mathcal{H}_d \in \mathbb{R}^{d\times K}$, $\mathcal{H}_D \in \mathbb{R}^{K\times d}$, $(\cdot)^{\mathrm{T}}$ denotes the transpose of the matrix. It is generally assumed that operations such as the activation function, dropout, feature upsampling and downsampling in MLP can help explore channel correlation in the high-dimensional features. Similarly, the transposition of channel and instance can also force IOA-MLP to learn the instance order correlation implicitly in $\mathcal{H}_t^L$.

Instance Distribution Aware Task. Unlike the instance relative order information implied in the Top-K key instances selection, the pseudo-labels can directly reflect the distribution of positive instances in the bag. Therefore, we design an *Instance Distribution Aware Task* (IDA-Task) with the following advantages. **1)** As an auxiliary task of predicting positive instances distribution, it can guide the bag-level feature with better global attention; **2)** It can be the regulation for MIL training to reduce overfitting, which helps to achieve higher performance in WSI-level tasks.

Specifically, we will assign labels to each WSI according to the positive distribution in the Top-K key instances. The proportion of positive instances in the selected Top-K are divided into four non-overlapping bins: $[r_0, r_1), [r_1, r_2), [r_2, r_3), [r_3, r_4)$. Simply, we set $r_0 = 0$, $r_1 = 0.25$, $r_2 = 0.5$, $r_3 = 0.75$, $r_4 = +\infty$. For instance-level distribution labels Y_I, we can get it as follows:

$$Y_I = i \text{ iff } Y_{I,ratio} \in [r_i, r_{i+1}). \tag{4}$$

Given the bag-level feature $\mathcal{H}_B \in \mathbb{R}^d$, the loss function of the instance distribution aware task and bag-level prediction task can be defined as follows:

$$\begin{aligned} L_{\text{instance}} &= \mathrm{L_I}\left(Y_I, \operatorname{softmax}\left(\mathcal{H}_B\right)\right), \\ L_{\text{bag}} &= \mathrm{L_B}\left(Y, \operatorname{softmax}\left(\mathcal{H}_B\right)\right), \\ L_{\text{total}} &= L_{\text{bag}} + \lambda L_{\text{instance}}, \end{aligned} \tag{5}$$

where $\mathrm{L_I}$ is the cross entropy loss function, $\mathrm{L_B}$ is the loss function for a specific bag-level prediction task. λ denotes the intensity of instance distribution aware.

3.3.2 Weakening Bag Semantical Unalignment with Bag-level Semantically Guided Attention

Since not all patches can inherit WSI-level labels, *e.g.*, for a tumor WSI, the tumor patches may be less than 10%. Such semantical unalignment between bag-level labels and instances still inevitably exists in Top-K key instances. Intuitively, instances with less semantical unalignment to the bag-level label should have lower weights. Therefore, we design a *Bag-level Semantically Guided Attention* (BG-Attn) to reduce the weight of semantical unaligned patches and strengthen the weight of semantical aligned patches. Given the output of Transformer encoder $\mathcal{H}_D \in \mathbb{R}^{K\times d}$, the procedure can be defined as follows:

$$\begin{aligned} \alpha_i &= \frac{\sum_{c=1}^{C} \exp\left(\boldsymbol{h}_i^D \boldsymbol{w}_c + b_c\right)}{\sum_{k=1}^{K}\sum_{c=1}^{C} \exp\left(\boldsymbol{h}_k^D \boldsymbol{w}_c + b_c\right)}, \\ \mathcal{H}_G &= \operatorname{Concat}\left(\alpha_1 \boldsymbol{h}_1^D, \ldots, \alpha_K \boldsymbol{h}_K^D\right), \end{aligned} \tag{6}$$

where $\boldsymbol{h}_i^D \in \mathbb{R}^d$, $\boldsymbol{w}_c \in \mathbb{R}^{d\times 1}$ is the c-th column vector of $\boldsymbol{W}_c \in \mathbb{R}^{d\times C}$, b_c is a bias in $\boldsymbol{b}_c \in \mathbb{R}^C$, $\mathcal{H}_G \in \mathbb{R}^{K\times d}$, C is bag-level category, K is the number of key instances.

Datasets	Total	FSL			WSL		
		0.1% / 0.5%	1%	Test	Train	Val	Test
Camelyon16	399	270 (with LPA)			216	54	129
	600K	2.8K	5.6K	40K		/	
CRC-Surv	444	/			267	66	111
	100K	0.08K	0.8K	20K		/	

Table 1. **Dataset Description.** We report the total number of WSIs and patches used in each dataset. **FSL.** Instance-level fully supervised learning with LPA, *i.e.* 0.1%/0.5% and 1% labeled patches. The performance of the LPA-trained weak classifier on the test set is shown in the *Supplement*. **WSL.** Bag-level weakly supervised learning with WA. **Camelyon16.** Camelyon16 [5] contains *both patch and WSI-level annotations*. It includes 270 WSIs in the training set and 129 WSIs in the test set (*officially splitting*). 600K patches (tumor: 300K, normal: 300K) are selected for the FSL. **CRC-Surv.** We adopt the TCGA-COAD [32] (444 WSIs, only contains *WSI-level annotations*) and the NCT-CRC-HE [23] (100K patches, only contains *patch-level annotations*).

Inspired by [42], we use a bag-level MLP to predict the score of each instance, *i.e.*, each instance is considered as a bag containing only a single instance to measure the alignment with bag-level semantical information. Lower attention weight α_i will be given to the instance that has less semantical alignment to bag-level label.

4. Experiments

Downstream Tasks. To verify the effectiveness of our proposed LNPL-MIL framework, as shown in Tab. 1, we conduct experiments on two representative downstream tasks, including the Tumor Diagnosis (Camelyon16) and Survival Prediction (CRC-Surv). Besides, we will provide discussions in two proportions of LPA, *i.e.*, 0.1%/0.5% and 1% labeled patch-level annotations.

Implementation Details. We use 4-fold cross-validation for all experiments and report the results of all models in the form of mean_{std}. We bold the **best** and underline the <u>second best</u>. Besides, for the results in survival prediction, "†" denotes P-Value<0.05. After filtering the background area, the WSI is split into a series of 224×224 sized patches. Among them, the Camelyon16 dataset is processed at 40×, on average, each WSI includes 30,068 patches. TCGA-COAD is processed at 20×, on average, each WSI includes 13,414 patches. For the model's architecture and training parameters, we use a 4-layer Transformer, 1D convolution with size 3, trained with the Ranger optimizer [52]. The learning rate is 2e-4, and the batch size is 1. The discrete position encoding method mentioned in [7] is adopted in Transformer. For bag-level loss functions, we use cross-entropy loss for Tumor Diagnosis and cross entropy-based

	0.5% Labeled			1% Labeled		
t_{ROI}	100	200	400	100	200	400
w/o	0.410	0.406	0.413	0.489	0.501	0.508
SSL	0.537	0.558	0.570	0.550	<u>0.574</u>	0.586
0.2	0.537	0.558	0.570	<u>0.556</u>	0.567	0.586
0.4	<u>0.541</u>	<u>0.566</u>	<u>0.578</u>	0.552	0.566	<u>0.590</u>
0.6	**0.546**	**0.571**	**0.584**	**0.574**	**0.603**	**0.616**

Table 2. **Performance of the SP-LNPL method in Patch-level Tumor Region Detection.** Comparison of the tumor region detection ability. **Top Row.** *w/o.* We compare with the LPA-trained classifier without SP-LNPL method. *SSL.* We select SOTA SSL method PAWS [3] for comparison. **Bottom Row.** *"0.x"*. It's the ratio threshold of positive ROI pseudo-labels $\hat{y}_{roi}$. We explore the effects of ROI super patches threshold t_{ROI} in the SP-LNPL. **FROC.** The FROC metric we report is the average sensitivity of 5 false positive rates: 1/2, 1, 2, 4, and 8 FPs per WSI. The results are reported on the Camelyon16 test set. A higher FROC metric represents better performance.

Cox proportional loss function for Survival Prediction following [11, 61]. For the SP-LNPL, we set the size of the super patch to 50. The ratio threshold of positive instances in each super patch t_{ROI} is 10/20, and the selection of Top-K key instances is 400/200 with (0.1%/0.5%)/1% annotation. For the setting of λ in IDA-Task, the Camelyon16 dataset is set to 0.001/0.001 with 0.5%/1% annotation, and the TCGA-COAD dataset is set to 0.001/0.01 with 0.1%/1% annotation. The setting of hyperparameters is discussed further in the ablation study and *Supplement*.

4.1. Experiments on Tumor Diagnosis

Patch-level Tumor Region Detection. We evaluate the performance of the SP-LNPL method over the Camelyon16 dataset, which has pixel-level annotations. Therefore, the patch-level tumor region detection ability of selected Top-K key instances can be compared. As shown in Tab. 2, we have the following observations: **1)** Take the performance of FSL-trained classifier as the baseline, both the SSL and SP-LNPL methods can significantly reduce the false positives in the selected Top-K key instances. **2)** Classifiers trained with more labeled data generally have higher FROC. Besides, a larger t_{ROI} can also help the SP-LNPL to get a higher FROC. The selection of t_{ROI} will be discussed further in the ablation study.

Weakly Supervised Comparison. Tumor Diagnosis results are summarized in Tab. 3. We first train a weak classifier with 0.5% or 1% labeled patches. Then we employ the weak classifier to select Top-K key instances of high tumor probabilities for each WSI in Camelyon16 [5]. We have the following observations: **1)** The tumor area in the Camelyon16 is generally small (less than 10%). The random

	Tumor Diagnosis		
Architecture	0% Labeled	0.5% Labeled	1% Labeled
AB-MIL [34]	$0.840_{.024}$	$0.877_{.030}$	$0.873_{.006}$
CLAM-SB [34]	$0.819_{.043}$	$0.880_{.020}$	$0.877_{.010}$
Deep-Attn [60]	$0.536_{.116}$	$0.863_{.014}$	$0.841_{.010}$
Loss-Attn [42]	$0.857_{.023}$	$0.900_{.035}$	$0.910_{.012}$
DTFD-MIL [62]	$0.844_{.046}$	$0.909_{.008}$	$0.902_{.033}$
DTFD-MIL [62]	0.946[1]	/	/
DS-MIL [28]	$0.743_{.066}$	$0.821_{.058}$	$0.792_{.039}$
FR-MIL [16]	$\underline{0.898}_{.066}$	$0.914_{.017}$	$0.902_{.017}$
GCN-MIL [31]	$0.896_{.032}$	$0.903_{.009}$	$0.943_{.009}$
Patch-GCN [11]	$\mathbf{0.925}_{.020}$	$\underline{0.944}_{.005}$	$\underline{0.957}_{.003}$
Mixed-Trans [7]	/	$0.746_{.028}$	$0.755_{.027}$
LNPL-MIL (Ours)	/	$\mathbf{0.971}_{.011}$	$\mathbf{0.986}_{.007}$

[1] The paper reports the result using ImageNet pre-trained ResNet-50 features under the magnification of 20×.

Table 3. **Tumor Diagnosis.** Comparison of AUC performance in Camelyon16. **Top Row.** Bypass attention based MIL model [34, 42, 60, 62]. **Second Row.** Non-local attention [28] and self-attention based [16] MIL model. **Third Row.** GNN based MIL model [11, 16]. **Bottom Row.** Transformer based MIL model with mixed supervision strategy [7]. We compare the experimental setup from three perspectives: **0% Labeled.** Only slide-level labels (all the patches); **0.5% Labeled.** Weak classifiers trained with 0.5% labeled data (Top-K patches); **1% Labeled.** Weak classifiers trained with 1% labeled data (Top-K patches).

sampling-based method Deep-Attn and coarse superpixels-based method Mixed-Trans cannot achieve good results. Top-K key instances selection is an efficient way to this problem. Since the WSI-level labels in Camelyon16 are only associated with the suspected tumor patches, when LPA are available, the classification results of all weakly supervised methods have been improved after Top-K key instances selection. **2)** Self-attention-based method FR-MIL and GNN-based method Patch-GCN benefit from the strong ability of instances correlation aggregation, and good results can be achieved. Since the SP-LNPL method can dramatically reduce false positives in selected Top-K and TOD-MIL can fully explore the instance order and distribution within the bag, the LNPL-MIL framework achieves at least 2.7% and 2.9% AUC improvement over a range of competing methods, with 0.5% and 1% Labeled, respectively.

4.2. Experiments on Survival Prediction

Survival prediction results are summarized in Tab. 4. For the 0.1% or 1% labeled data in the NCT-CRC-HE [23], we first train a weak classifier for nine tissue classifications. Then we follow the tissue types selected in [1]. For each WSI in TCGA-COAD [32], we select Top-K key instances that belong to lymphocytes, cancer-associated stroma, or

Architecture	Survival Prediction 0% Labeled	0.1% Labeled	1% Labeled
AB-MIL [34]	$0.582_{.069}$	$\underline{0.601}^{\dagger}_{.057}$	$0.592_{.068}$
Deep-Attn [60]	$0.557_{.076}$	$0.558_{.089}$	$0.561_{.076}$
Loss-Attn [42]	$0.556_{.074}$	$0.553_{.072}$	$0.555_{.067}$
DS-MIL [28]	$0.564_{.068}$	$0.552_{.060}$	$0.540_{.076}$
GCN-MIL [31]	$\mathbf{0.588}_{.066}$	$0.593_{.040}$	$0.574^{\dagger}_{.054}$
Patch-GCN [11]	$\underline{0.580}_{.024}$	$0.578^{\dagger}_{.022}$	$\underline{0.598}_{.042}$
Mixed-Trans [7]	/	$0.547_{.069}$	$0.533_{.066}$
LNPL-MIL (Ours)	/	$\mathbf{0.627}^{\dagger}_{.043}$	$\mathbf{0.621}^{\dagger}_{.074}$

Table 4. **Survival Prediction.** Comparison of C-Index performance in CRC-Surv. The specific experimental setup is the same as Tumor Diagnosis. It should be noted that some of the methods are *not applicable* to the Survival Prediction, so we do not include them in the comparison experiment.

colorectal adenocarcinoma epithelium. We have the following observations: **1)** In CRC, a highly heterogeneous cancer [10, 44, 53], most methods cannot achieve satisfactory results at 0% Labeled (all the patches). Even worse, when the two forms of annotation come from different centers, the noisy pseudo-labels of the weak classifier will be amplified. Therefore, based on the Top-K key instances selected without data cleaning, many MIL methods cannot achieve better results due to the impact of false positives. **2)** We find that Top-K key instances selected with the 1% Labeled are not always more suitable than 0.1% Labeled in the Survival Prediction. We guess that since the hazard of patients in Survival Prediction is often related to the tumor microenvironment, consisting of many patches rather than a single tumor patch, the accuracy of key instances is not the only determinant of the Survival Prediction, *e.g.*, the spatial correlation of instances in the bag is also an important factor. **3)** Subject to the difficulties, the LNPL-MIL framework relies on more robust Top-K key instances selection and stronger correlation aware ability. It can still achieve at least 2.6% and 2.3% C-Index improvement over a range of competing methods, with 0.1% and 1% Labeled, respectively.

4.3. Ablation Study

Effects of Different Settings in SP-LNPL. Ablation results are summarized in Tab. 5. We have following observations: **1)** Similar conclusions as in Tab. 1: the FSL assisted with SP-LNPL in WSI-level tasks can be better than FSL and SSL. We further demonstrate the effect of several representative MIL methods assisted with SP-LNPL in the *Supplement*, and the performance of MIL methods can be better. **2)** Although a higher t_{ROI} can achieve better results on the FROC metric, it may also lead to missing key instances, so we suggest choosing a lower t_{ROI} like 0.2 or

t_{ROI}	Tumor Diagnosis 0.5%	1%	Survival Prediction 0.1%	1%
w/o	$0.902_{.040}$	$0.944_{.007}$	$0.625^{\dagger}_{.040}$	$0.606^{\dagger}_{.085}$
SSL	$0.945_{.015}$	$0.964_{.013}$	$0.589_{.013}$	$\underline{0.617}^{\dagger}_{.082}$
0.2	$\mathbf{0.971}_{.011}$	$\underline{0.980}_{.009}$	$\underline{0.627}^{\dagger}_{.043}$	$0.612_{.072}$
0.4	$\underline{0.965}_{.008}$	$\mathbf{0.986}_{.007}$	$\mathbf{0.630}^{\dagger}_{.045}$	$\mathbf{0.621}^{\dagger}_{.074}$
0.6	$0.945_{.004}$	$0.957_{.009}$	$0.601^{\dagger}_{.025}$	$0.554_{.027}$

Table 5. **Effects of Different Settings in SP-LNPL. Top Row.** We compare the performance of selected Top-K key instances based on FSL (w/o) and SOTA SSL method PAWS [3] in WSI-level label prediction. **Bottom Row.** *"0.x"*. It is the ratio threshold of positive ROI pseudo-labels. We explore the effects of positive ROI pseudo-labels threshold t_{ROI} in the SP-LNPL method for the Tumor Diagnosis and Survival Prediction.

Architecture	Tumor Diagnosis 0.5%	1%	Survival Prediction 0.1%	1%
w/o C-Trans	$0.926_{.015}$	$0.967_{.014}$	$0.584_{.032}$	$0.600^{\dagger}_{.073}$
w/o IOA-MLP	$0.950_{.011}$	$0.958_{.013}$	$0.616^{\dagger}_{.043}$	$\underline{0.614}^{\dagger}_{.069}$
w/ Shuffle	$0.930_{.015}$	$0.931_{.020}$	$0.597_{.018}$	$0.593_{.022}$
ℓ^1 norm	$0.938_{.024}$	$0.938_{.054}$	$0.567^{\dagger}_{.027}$	$0.596^{\dagger}_{.050}$
ℓ^2 norm	$0.943_{.041}$	$\underline{0.985}_{.008}$	$0.593^{\dagger}_{.019}$	$0.562^{\dagger}_{.049}$
w/o IDA-Task	$0.964_{.006}$	$0.983_{.008}$	$0.603^{\dagger}_{.027}$	$\underline{0.614}^{\dagger}_{.074}$
w/o BG-Attn	$0.948_{.021}$	$0.968_{.011}$	$\mathbf{0.629}^{\dagger}_{.041}$	$0.611^{\dagger}_{.064}$
w/ AB-MIL	$\underline{0.967}_{.009}$	$0.982_{.009}$	$0.611^{\dagger}_{.029}$	$0.608^{\dagger}_{.034}$
TOD-MIL	$\mathbf{0.971}_{.011}$	$\mathbf{0.986}_{.007}$	$\underline{0.627}^{\dagger}_{.043}$	$\mathbf{0.621}^{\dagger}_{.074}$

Table 6. **Effects of Different Modules in TOD-MIL. Top Row.** Effects of local and global communications. **Second Row.** Effects of instance order aware. *w/ Shuffle.* We report the $\mathrm{mean}_{\mathrm{std}}$ of the results for the shuffle method under five seeds. Specifically, we randomly shuffle the Top-K key instances at each training epoch. **Third Row.** Effects of instance distribution aware. *ℓ^1/ℓ^2 norm.* We compare with conventional regularization methods. Regularization constraints are added to the bag-level loss with a commonly used weight of 0.001. **Fourth Row.** Effects of bag-level semantically guided attention. *w/ AB-MIL.* We replace BG-Attn with AB-MIL. **Bottom Row.** Our proposed TOD-MIL in LNPL-MIL.

0.4. In the *Supplement*, we discuss the effect of super patch size on the SP-LNPL and find a medium size like 50 works better. Besides, we also discuss the influence of the proportion for labeled patches on the parameter selection. We find weak classifiers trained with fewer annotations must be more conservative in selecting relevant parameters.

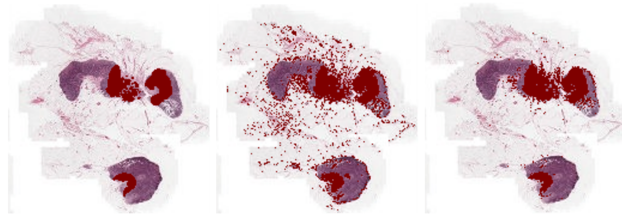
Effects of Different Modules in TOD-MIL. Ablation results are summarized in Tab. 6. We have following observations: **1)** In the top row, we find that since the C-Trans can explore local-global correlation information, it is an essential foundation in the TOD-MIL. **2)** IOA-MLP learns positive probability order information from LPA-trained weak classifiers and improves the MIL performance. Besides, during the MIL training, the random shuffle operation will damage the order of the Top-K key instances in each epoch, forcing the model to keep fitting the wrong order information. Results show that it will lead to poor performance. **3)** IDA-Task improves the performance of the bag-level task by introducing instance distribution labels as supervision. Besides, it can also have higher promotion than conventional regularization such as ℓ^1/ℓ^2 norm. **4)** Both BG-Attn and AB-MIL are based on the attention mechanism. The difference is that BG-Attn uses the bag-level classifier for attention calculation. Since BG-Attn introduces bag-level semantical supervision to the attention calculation, it can achieve better results than AB-MIL. We also notice a negative optimization under the 0.1% annotation of the CRC. We guess that since the nine categories classifier trained under the 0.1% labeled data (80 patches) has lower accuracy, much unrelated tissue types are introduced into Top-K key instances. It causes the attention weight obtained by BG-Attn to be relatively similar and cannot effectively weaken unrelated patches.

4.4. Visualization Analysis

For the SP-LNPL, we perform visual analysis from both global and local aspects, and the visualization results are shown in Fig. 3 and Fig. 4, respectively. We have the following observations: **1)** In Fig. 3, compared the ground-truth (a) with the prediction (b) and (c), the SP-LNPL method can greatly reduce the false positives and help weak classifiers select more accurate Top-K key instances. **2)** In Fig. 4 (a), adopting KNN in the feature space can effectively cluster similar patches into the same super patch. Importantly, combining the pseudo-labels of the LPA-trained weak classifier with the result of ROI super patches can jointly remove false positives in Top-K key instances. In Fig. 4 (b), we find that due to weak generalization of the LPA-trained classifier, some similar or blurry corrupted patches will be assigned positive labels with high probability, resulting in the false positives during the Top-K key instances selection. More visualization results are in the *Supplement*.

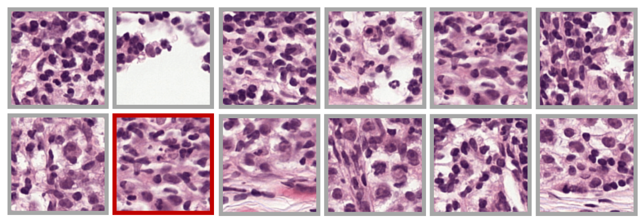
5. Conclusion

MIL is widely used in WSIs related tasks when only WA exists. As one of the weak annotation forms in WSIs, LPA are sometimes available in many tasks. Intuitively, assigning pseudo-labels to unlabeled data by LPA can promote MIL. However, the unreliable pseudo-labels will inevitably introduce noise. Currently, how to fully use LPA to promote MIL is still lack of exploration. In this paper, we design a framework called LNPL-MIL that learns from noisy pseudo labels to promote multiple instance learning.

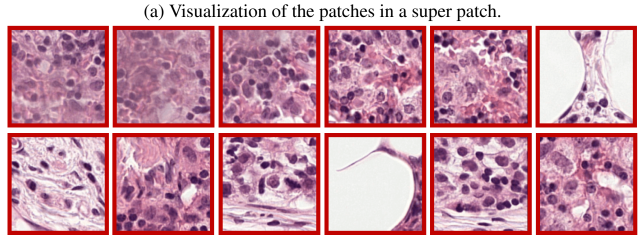


(a) Tumor 011. (b) Weak classifier. (c) Ours, $t_{\text{ROI}} = 0.6$.

Figure 3. **The Visualization of Weak Classifier Predictions (Camelyon16). (a).** Dark red represents the pathologist-annotated tumor area. **(b).** Dark red represents the tumor region predicted by the weak classifier after only FSL training. **(c).** Dark red represents the tumor region predicted by the weak classifier after FSL training and data cleaning by the SP-LNPL.



(a) Visualization of the patches in a super patch.



(b) Visualization of false positive patches in Top-K key instances: with a high probability of positive but out of the ROI super patch.

Figure 4. **The Visualization of Super Patch and False Positive Patches in Top-K Key Instances (Camelyon16).** *Red/grey* indicates that the weak classifier predicts the patch has a *high/low* probability of being positive. **(a).** We visualize the super patch obtained by the KNN search in the feature space. **(b).** We visualize the false positives filtered by the SP-LNPL method in Top-K key instances selection.

Specifically, for LPA-trained weak classifier, we design the SP-LNPL method to select more accurate Top-K key instances. Then, we propose the TOD-MIL that fully utilize instance order and distribution and weaken semantical unalignment in the MIL. We verify the LNPL-MIL framework on two typical WSI-related downstream tasks and achieve the state-of-the-art performance. Importantly, the improvement of 2.7%/2.9% in AUC and 2.6%/2.3% in C-Index can be achieved with the patches labeled for two scale for the Tumor Diagnosis and Survival Prediction, respectively. Besides, we discuss the effectiveness of proposed SP-LNPL and TOD-MIL in the ablation study. Visualization analysis further verifies the effectiveness. In the future, we will explore the potential for the combination of SSL in the LNPL, and verify the proposed LNPL-MIL framework in more tasks.

6. Acknowledgement

Haoqian Wang was funded through National Key Research and Development Program of China (Project No. 2022YFB36066), in part by the Shenzhen Science and Technology Project under Grant (JCYJ20220818101001004). Yongbing Zhang was supported in part by the National Natural Science Foundation of China (62031023), in part by the Shenzhen Science and Technology Project (JCYJ20200109142808034&GXWD20220818170353009), and in part by Guangdong Special Support (2019TX05X187).

References

[1] Christian Abbet, Inti Zlobec, Behzad Bozorgtabar, and Jean-Philippe Thiran. Divide-and-rule: self-supervised learning for survival analysis in colorectal cancer. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 480–489. Springer, 2020. 1, 2, 6

[2] Valentin Anklin, Pushpak Pati, Guillaume Jaume, Behzad Bozorgtabar, Antonio Foncubierta-Rodriguez, Jean-Philippe Thiran, Mathilde Sibony, Maria Gabrani, and Orcun Goksel. Learning whole-slide segmentation from inexact and incomplete labels using tissue graphs. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 636–646. Springer, 2021. 2, 4

[3] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Armand Joulin, Nicolas Ballas, and Michael Rabbat. Semi-supervised learning of visual features by non-parametrically predicting view assignments with support samples. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8443–8452, 2021. 6, 7

[4] Dara Bahri, Heinrich Jiang, and Maya Gupta. Deep k-nn for noisy labels. In *International Conference on Machine Learning*, pages 540–550. PMLR, 2020. 2

[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. 2, 5, 6

[6] Kaustav Bera, Kurt A Schalper, David L Rimm, Vamsidhar Velcheti, and Anant Madabhushi. Artificial intelligence in digital pathology—new tools for diagnosis and precision oncology. *Nature reviews Clinical oncology*, 16(11):703–715, 2019. 2

[7] Hao Bian, Zhuchen Shao, Yang Chen, Yifeng Wang, Haoqian Wang, Jian Zhang, and Yongbing Zhang. Multiple instance learning with mixed supervision in gleason grading. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 204–213. Springer, 2022. 2, 4, 5, 6, 7

[8] 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*, pages 1301–1309, 2019. 1, 2

[9] Marc-André Carbonneau, Veronika Cheplygina, Eric Granger, and Ghyslain Gagnon. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition*, 77:329–353, 2018. 2

[10] Richard J Chen, Chengkuan Chen, Yicong Li, Tiffany Y Chen, Andrew D Trister, Rahul G Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16144–16155, 2022. 2, 7

[11] Richard J Chen, Ming Y Lu, Muhammad Shaban, Chengkuan Chen, Tiffany Y Chen, Drew FK Williamson, and Faisal Mahmood. Whole slide images are 2d point clouds: Context-aware survival prediction using patch-based graph convolutional networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 339–349. Springer, 2021. 2, 6, 7

[12] Richard J Chen, Ming Y Lu, Wei-Hung Weng, Tiffany Y Chen, Drew FK Williamson, Trevor Manz, Maha Shady, and Faisal Mahmood. Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4025, 2021. 2

[13] Richard J Chen, Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Jana Lipkova, Muhammad Shaban, Maha Shady, Mane Williams, Bumjin Joo, Zahra Noor, et al. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. *Cancer Cell*, 2022. 1

[14] Yang Chen, Zhuchen Shao, Hao Bian, Zijie Fang, Yifeng Wang, Yuanhao Cai, Haoqian Wang, Guojun Liu, Xi Li, and Yongbing Zhang. dmil-transformer: Multiple instance learning via integrating morphological and spatial information for lymph node metastasis classification. *IEEE Journal of Biomedical and Health Informatics*, 2023. 2

[15] P. Chikontwe, Meejeong Kim, S. Nam, H. Go, and S. Park. Multiple instance learning with center embeddings for histopathology classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 519–528, 2020. 2

[16] Philip Chikontwe, Soo Jeong Nam, Heounjeong Go, Meejeong Kim, Hyun Jung Sung, and Sang Hyun Park. Feature re-calibration based multiple instance learning for whole slide image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 420–430. Springer, 2022. 2, 6

[17] Zeyu Gao, Bangyang Hong, Yang Li, Xianli Zhang, Jialun Wu, Chunbao Wang, Xiangrong Zhang, Tieliang Gong,

Yefeng Zheng, Deyu Meng, et al. A semi-supervised multi-task learning framework for cancer classification with weak annotation in whole-slide images. *Medical Image Analysis*, page 102652, 2022. 1, 2

[18] Zeyu Gao, Pargorn Puttapirat, Jiangbo Shi, and Chen Li. Renal cell carcinoma detection and subtyping with minimal point-based annotation in whole-slide images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 439–448. Springer, 2020. 1, 2

[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. 3

[20] Dan Hendrycks, Mantas Mazeika, Duncan Wilson, and Kevin Gimpel. Using trusted data to train deep networks on labels corrupted by severe noise. *Advances in neural information processing systems*, 31, 2018. 2

[21] Wentai Hou, Lequan Yu, Chengxuan Lin, Helong Huang, Rongshan Yu, Jing Qin, and Liansheng Wang. H2-mil: Exploring hierarchical representation with heterogeneous multiple instance learning for whole slide image analysis. 2022. 2

[22] Ziwang Huang, Hua Chai, Ruoqi Wang, Haitao Wang, Yuedong Yang, and Hejun Wu. Integration of patch features through self-supervised learning and transformer for survival analysis on whole slide images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 561–570. Springer, 2021. 2

[23] Jakob Nikolas Kather, Niels Halama, and Alexander Marx. 100,000 histological images of human colorectal cancer and healthy tissue, Apr. 2018. 5, 6

[24] Jakob Nikolas Kather, Alexander T Pearson, Niels Halama, Dirk Jäger, Jeremias Krause, Sven H Loosen, Alexander Marx, Peter Boor, Frank Tacke, Ulf Peter Neumann, et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nature medicine*, 25(7):1054–1056, 2019. 2

[25] Andreas Kleppe, Ole-Johan Skrede, Sepp De Raedt, Knut Liestøl, David J Kerr, and Håvard E Danielsen. Designing deep learning studies in cancer diagnostics. *Nature Reviews Cancer*, 21(3):199–211, 2021. 2

[26] Daisuke Komura and Shumpei Ishikawa. Machine learning methods for histopathological image analysis. *Computational and structural biotechnology journal*, 16:34–42, 2018. 1

[27] Marvin Lerousseau, Maria Vakalopoulou, Marion Classe, Julien Adam, Enzo Battistella, Alexandre Carré, Théo Estienne, Théophraste Henry, Eric Deutsch, and Nikos Paragios. Weakly supervised multiple instance learning histopathological tumor segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 470–479, 2020. 2

[28] 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 Conference on Computer Vision and Pattern Recognition*, 2021. 2, 6, 7

[29] Hang Li, Fan Yang, Yu Zhao, Xiaohan Xing, Jun Zhang, Mingxuan Gao, Junzhou Huang, Liansheng Wang, and Jianhua Yao. Dt-mil: Deformable transformer for multi-instance learning on histopathological image. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 206–216. Springer, 2021. 2

[30] Jiahui Li, Shuang Yang, Xiaodi Huang, Qian Da, Xiaoqun Yang, Zhiqiang Hu, Qi Duan, Chaofu Wang, and Hongsheng Li. Signet ring cell detection with a semi-supervised learning framework. In *International conference on information processing in medical imaging*, pages 842–854. Springer, 2019. 2

[31] Ruoyu Li, Jiawen Yao, Xinliang Zhu, Yeqing Li, and Junzhou Huang. Graph cnn for survival analysis on whole slide pathological images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 174–182. Springer, 2018. 6, 7

[32] Jianfang Liu, Tara Lichtenberg, Katherine A Hoadley, Laila M Poisson, Alexander J Lazar, Andrew D Cherniack, Albert J Kovatich, Christopher C Benz, Douglas A Levine, Adrian V Lee, et al. An integrated tcga pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell*, 173(2):400–416, 2018. 2, 5, 6

[33] Ming Y Lu, Tiffany Y Chen, Drew FK Williamson, Melissa Zhao, Maha Shady, Jana Lipkova, and Faisal Mahmood. Ai-based pathology predicts origins for cancers of unknown primary. *Nature*, 594(7861):106–110, 2021. 1, 2

[34] 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. 1, 2, 6, 7

[35] Yu A Malkov and Dmitry A Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine intelligence*, 42(4):824–836, 2018. 4

[36] Niccolo Marini, Sebastian Otálora, Henning Müller, and Manfredo Atzori. Semi-supervised training of deep convolutional neural networks with heterogeneous data and few local annotations: An experiment on prostate histopathology image classification. *Medical image analysis*, 73:102165, 2021. 1

[37] Sachin Mehta, Ximing Lu, Wenjun Wu, Donald Weaver, Hannaneh Hajishirzi, Joann G Elmore, and Linda G Shapiro. End-to-end diagnosis of breast biopsy images with transformers. *Medical Image Analysis*, 79:102466, 2022. 2

[38] Linhao Qu, Xiaoyuan Luo, Manning Wang, and Zhijian Song. Bi-directional weakly supervised knowledge distillation for whole slide image classification. *arXiv preprint arXiv:2210.03664*, 2022. 2

[39] Benoît Schmauch, Alberto Romagnoni, Elodie Pronier, Charlie Saillard, Pascale Maillé, Julien Calderaro, Aurélie Kamoun, Meriem Sefta, Sylvain Toldo, Mikhail Zaslavskiy, et al. A deep learning model to predict rna-seq expression of tumours from whole slide images. *Nature communications*, 11(1):1–15, 2020. 1

[40] 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. 2

[41] Zhuchen Shao, Yang Chen, Hao Bian, Jian Zhang, Guojun Liu, and Yongbing Zhang. Hvtsurv: Hierarchical vision transformer for patient-level survival prediction from whole slide image. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2209–2217, 2023. 2

[42] Xiaoshuang Shi, Fuyong Xing, Yuanpu Xie, Zizhao Zhang, Lei Cui, and Lin Yang. Loss-based attention for deep multiple instance learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5742–5749, 2020. 5, 6, 7

[43] Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-weight-net: Learning an explicit mapping for sample weighting. *Advances in neural information processing systems*, 32, 2019. 2

[44] Ole-Johan Skrede, Sepp De Raedt, Andreas Kleppe, Tarjei S Hveem, Knut Liestøl, John Maddison, Hanne A Askautrud, Manohar Pradhan, John Arne Nesheim, Fritz Albregtsen, et al. Deep learning for prediction of colorectal cancer outcome: a discovery and validation study. *The Lancet*, 395(10221):350–360, 2020. 7

[45] Hwanjun Song, Minseok Kim, and Jae-Gil Lee. Selfie: Refurbishing unclean samples for robust deep learning. In *International Conference on Machine Learning*, pages 5907–5915. PMLR, 2019. 2

[46] Chetan L Srinidhi, Ozan Ciga, and Anne L Martel. Deep neural network models for computational histopathology: A survey. *Medical Image Analysis*, 67:101813, 2021. 1

[47] Chetan L Srinidhi, Seung Wook Kim, Fu-Der Chen, and Anne L Martel. Self-supervised driven consistency training for annotation efficient histopathology image analysis. *Medical Image Analysis*, 75:102256, 2022. 4

[48] Ryutaro Tanno, Ardavan Saeedi, Swami Sankaranarayanan, Daniel C Alexander, and Nathan Silberman. Learning from noisy labels by regularized estimation of annotator confusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11244–11253, 2019. 2

[49] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in Neural Information Processing Systems*, 34:24261–24272, 2021. 4

[50] Yisen Wang, Weiyang Liu, Xingjun Ma, James Bailey, Hongyuan Zha, Le Song, and Shu-Tao Xia. Iterative learning with open-set noisy labels. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8688–8696, 2018. 2

[51] Zhenzhen Wang, Aleksander S Popel, and Jeremias Sulam. Label cleaning multiple instance learning: Refining coarse annotations on single whole-slide images. *arXiv preprint arXiv:2109.10778*, 2021. 2

[52] Less Wright. Ranger - a synergistic optimizer. https://github.com/lessw2020/Ranger-Deep-Learning-Optimizer, 2019. 5

[53] Ellery Wulczyn, David F Steiner, Melissa Moran, Markus Plass, Robert Reihs, Fraser Tan, Isabelle Flament-Auvigne, Trissia Brown, Peter Regitnig, Po-Hsuan Cameron Chen, et al. Interpretable survival prediction for colorectal cancer using deep learning. *NPJ digital medicine*, 4(1):1–13, 2021. 7

[54] Xiaobo Xia, Tongliang Liu, Bo Han, Chen Gong, Nannan Wang, Zongyuan Ge, and Yi Chang. Robust early-learning: Hindering the memorization of noisy labels. In *International conference on learning representations*, 2020. 2

[55] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2691–2699, 2015. 2

[56] Chensu Xie, Hassan Muhammad, Chad M Vanderbilt, Raul Caso, Dig Vijay Kumar Yarlagadda, Gabriele Campanella, and Thomas J Fuchs. Beyond classification: Whole slide tissue histopathology analysis by end-to-end part learning. In *Medical Imaging with Deep Learning*, pages 843–856. PMLR, 2020. 2

[57] G. Xu, Zhigang Song, Zhuo Sun, Calvin Ku, Z. Yang, C. Liu, S. Wang, Jianpeng Ma, and W. Xu. Camel: A weakly supervised learning framework for histopathology image segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10681–10690, 2019. 2

[58] Rikiya Yamashita, Jin Long, Teri Longacre, Lan Peng, Gerald Berry, Brock Martin, John Higgins, Daniel L Rubin, and Jeanne Shen. Deep learning model for the prediction of microsatellite instability in colorectal cancer: a diagnostic study. *The Lancet Oncology*, 22(1):132–141, 2021. 1

[59] Jiangchao Yao, Jiajie Wang, Ivor W Tsang, Ya Zhang, Jun Sun, Chengqi Zhang, and Rui Zhang. Deep learning from noisy image labels with quality embedding. *IEEE Transactions on Image Processing*, 28(4):1909–1922, 2018. 2

[60] Jiawen Yao, Xinliang Zhu, Jitendra Jonnagaddala, Nicholas Hawkins, and Junzhou Huang. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. *Medical Image Analysis*, 65:101789, 2020. 6, 7

[61] Shekoufeh Gorgi Zadeh and Matthias Schmid. Bias in cross-entropy-based training of deep survival networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(9):3126–3137, 2020. 6

[62] 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. 2, 6