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

Zhuchen Shao\textsuperscript{1}, Yifeng Wang\textsuperscript{2}, Yang Chen\textsuperscript{1}, Hao Bian\textsuperscript{1}, Shaohui Liu\textsuperscript{3}, Haoqian Wang\textsuperscript{1}, Yongbing Zhang\textsuperscript{2*}

\textsuperscript{1}Tsinghua Shenzhen International Graduate School, Tsinghua University
\textsuperscript{2}Harbin Institute of Technology (Shenzhen)\textsuperscript{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 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 × 50,000 pixels) of Whole Slide Images (WSIs) [46] and the biomedical backgrounds required for data annotations, 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 Supervised 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...
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 features 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 pseudolabels 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.
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 $X = \{x_1, y_1, \ldots, x_n, y_n\}$, 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 $A$ contains $Y$, and Dataset $B$ 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:

$$y_p = F_{\text{weak}}(x), \hat{y}_p = \arg\max_y y_p,$$

where $y_p$ denotes the positive probability, $\hat{y}_p$ denotes the assigned pseudo-label, $x$ denotes the input patch, and $F_{\text{weak}}$ 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. Then, the LPA-trained weak classifier is employed to give guidance to select key instances. 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 features \( H \). Then, we employ the \( H \) to perform global KNN search [35] to classify similar patches into the same super patch. For Problem 2), we first divide the \( H \) 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 for patches in different WSIs is distinct, we adopt the adaptive threshold method to reassign ROI pseudo-labels for patches in different WSIs.

Super-patch-based LNPL Method

Algorithm 1

**Input:** The bag \( X \) with a series of instances \( \{(x_1, y_{p,1}), \ldots, (x_n, y_{p,n})\} \). 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.
   \( H_0 \)
2. Feature pre-processing, padding for feature sequence \( H \).
   \( H_0, X_a \leftarrow \text{Padding} \left( (H, X) \right) \)
3. ROI pseudo label pre-processing.
   \( y_{roi} \leftarrow \text{Find the median positive probability} \)
for \( i \in [1, n] \) do \( \hat{y}_{roi,i} \leftarrow 1 \) if \( y_{roi,i} > y_{roi} \) else 0;
4. KNN search, looking for ROI super patches.
   Initialize \( X \) as \( \emptyset \)
for \( i \in [0 : N : w] \) do
   1. Select the \( w \) features closest to \( h_1 \) in \( H_a \).
      \( \text{idx} \leftarrow \text{Hsw. query} \left( h_1, \text{topn} = w \right) \)
   2. Count the ratio of positive ROI pseudo-labels.
      \( \text{ratio} \leftarrow \text{Count} \left( \hat{y}_{roi, \text{idx}} = 1 \right) \)
   3. Determine whether the cluster is an ROI super patch.
      if \( \text{ratio} > t_{roi} \) then \( X \leftarrow X + X_a \left( \text{idx} \right) \); \( X_a \leftarrow X_a - X_a \left( \text{idx} \right) \) \( \hat{y}_{roi, \text{idx}} \leftarrow 1 \) \( \text{idx} \) from \( X_a \)
5. Select Top-K key instances from the filtered bag \( X \).
   \( \hat{x}_1, \ldots, \hat{x}_K \leftarrow \text{Max} \left( \hat{y}_{p,1}, \ldots, \hat{y}_{p,n} \right) \)

**Algorithm 1 Super-patch-based LNPL Method**

**Problem**

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. 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 \( H_0^\ell \in \mathbb{R}^{K \times d} \), the procedure can be defined as follows:

\[
H_0^\ell = \text{Conv}(H_0^{\ell-1}) + H_0^{\ell-1}, \quad \ell = 1 \ldots L
\]

\[
H_0^\ell = \text{MSA}(\text{LN}(H_0^\ell)) + H_0^\ell, \quad \ell = 1 \ldots L
\]

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 \( H_0^L \in \mathbb{R}^{K \times d} \), the procedure can be defined as follows:

\[
H_d = \text{MLP}(H_0^T) + H_0^T,
\]

\[
H_D = \text{MLP}(H_D^T) + H_D^T,
\]

where \( H_d \in \mathbb{R}^{d \times K}, H_D \in \mathbb{R}^{K \times d}, (\cdot)^T \) 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 \( H_D^T \).
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: \([r_0, r_1), [r_1, r_2), [r_2, r_3), [r_3, +\infty)\). 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:

\[
Y_i = i \text{ if } Y_{i, \text{ratio}} \in [r_i, r_{i+1}).
\]

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

\[
L_{\text{instance}} = L_1 (Y_i, \text{softmax} (H_B)), \quad L_{\text{bag}} = L_B (Y, \text{softmax} (H_B)), \quad L_{\text{total}} = L_{\text{bag}} + \lambda L_{\text{instance}},
\]

where \(L_1\) is the cross entropy loss function, \(L_B\) is the loss function for a specific bag-level prediction task, \(\lambda\) 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 \(H_D \in \mathbb{R}^{K \times d}\), the procedure can be defined as follows:

\[
\alpha_i = \frac{\sum_{c=1}^{C} \exp \left( h_i^D w_c + b_c \right)}{\sum_{k=1}^{K} \sum_{c=1}^{C} \exp \left( h_k^D w_c + b_c \right)},
\]

\[
H_G = \text{Concat} \left( \alpha_1 h_1^D, \ldots, \alpha_K h_K^D \right),
\]

where \(h_i^D \in \mathbb{R}^d, w_c \in \mathbb{R}^{d \times 1}\) is the \(c\)-th column vector of \(W_c \in \mathbb{R}^{d \times C}\), \(b_c\) is a bias in \(b_c \in \mathbb{R}^C\), \(H_G \in \mathbb{R}^{K \times d}\), \(C\) is bag-level category, \(K\) is the number of key instances.

### 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\(_{\text{std}}\). We bold the best and underline the second best. 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

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Total</th>
<th>FSL 0.1%/0.5% 1% Test</th>
<th>WSL Train Val Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Camelyon16</td>
<td>600K</td>
<td>270 (with LPA) 2.8K 5.6K 40K</td>
<td>216 54 129</td>
</tr>
<tr>
<td>CRC-Surv</td>
<td>100K</td>
<td>0.8K 0.8K 20K</td>
<td>267 66 111</td>
</tr>
</tbody>
</table>

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 Suppement. 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).
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 $\lambda$ in $\lambda$-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 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
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 Table 5. We have the 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 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 Local and Global Communications

Effects of instance distribution aware. ${\ell^1, \ell^2}$ 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.

### Table 4. Survival Prediction

<table>
<thead>
<tr>
<th>Architecture</th>
<th>0% Labeled</th>
<th>0.1% Labeled</th>
<th>1% Labeled</th>
</tr>
</thead>
<tbody>
<tr>
<td>AB-MIL [34]</td>
<td>0.582_{0.09}</td>
<td>0.601_{0.07}</td>
<td>0.592_{0.06}</td>
</tr>
<tr>
<td>Deep-Attn [60]</td>
<td>0.557_{0.07}</td>
<td>0.558_{0.09}</td>
<td>0.561_{0.07}</td>
</tr>
<tr>
<td>Loss-Attn [42]</td>
<td>0.556_{0.047}</td>
<td>0.553_{0.072}</td>
<td>0.559_{0.067}</td>
</tr>
<tr>
<td>DS-MIL [28]</td>
<td>0.564_{0.068}</td>
<td>0.552_{0.060}</td>
<td>0.540_{0.076}</td>
</tr>
<tr>
<td>GCN-MIL [31]</td>
<td>0.588_{0.066}</td>
<td>0.593_{0.040}</td>
<td>0.574_{0.054}</td>
</tr>
<tr>
<td>Patch-GCN [11]</td>
<td>0.585_{0.024}</td>
<td>0.578_{0.022}</td>
<td>0.598_{0.042}</td>
</tr>
<tr>
<td>Mixed-Trans [7]</td>
<td>/</td>
<td>0.547_{0.069}</td>
<td>0.533_{0.066}</td>
</tr>
<tr>
<td>LNPL-MIL (Ours)</td>
<td>/</td>
<td>0.627_{0.043}</td>
<td>0.621_{0.074}</td>
</tr>
</tbody>
</table>

### 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. “@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.

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Tumor Diagnosis</th>
<th>Survival Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o C-Trans</td>
<td>0.926_{0.015}</td>
<td>0.967_{0.014}</td>
</tr>
<tr>
<td>w/ Shuffle</td>
<td>0.950_{0.011}</td>
<td>0.958_{0.013}</td>
</tr>
<tr>
<td>$\ell^1$ norm</td>
<td>0.932_{0.024}</td>
<td>0.938_{0.054}</td>
</tr>
<tr>
<td>$\ell^2$ norm</td>
<td>0.943_{0.041}</td>
<td>0.945_{0.008}</td>
</tr>
<tr>
<td>w/ IOA-MLP</td>
<td>0.964_{0.066}</td>
<td>0.983_{0.008}</td>
</tr>
<tr>
<td>w/ BA-MIL</td>
<td>0.948_{0.021}</td>
<td>0.968_{0.111}</td>
</tr>
<tr>
<td>w/ AB-MIL</td>
<td>0.967_{0.009}</td>
<td>0.982_{0.009}</td>
</tr>
<tr>
<td>TOD-MIL</td>
<td>0.971_{0.011}</td>
<td>0.986_{0.007}</td>
</tr>
</tbody>
</table>

### 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 mean 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. $\ell^1$ or $\ell^2$ 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.
tial 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 $\ell^1/\ell^2$ 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. 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 combina-
tion 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 (JCYJ2020010914280834&GXWD20220818170353009), and in part by Guangdong Special Support (2019TX05X187).

References


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


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


