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

1. Ablation Study

Effects of SP-LNPL for MIL Methods MIL methods results are summarized in Tab. 1. The SP-LNPL method can improve the performance of several representative MIL methods on two typical weakly supervised tasks. Besides, we have the following observation: **1)** In the Camelyon16-based Tumor Diagnosis, the correspondence between WSI-level labels and patches is relatively direct, so reducing false positives by SP-LNPL in Top-K key instances can significantly improve the performance. **2)** In the CRC-based Survival Prediction, the bag-level labels and the semantical information of a single instance are completely unaligned. In contrast, bag-level labels aligned with ROI regions such as the tumor microenvironment. The instances obtained by Top-K selection without SP-LNPL have many false positives and may be more spatially discrete, resulting in suboptimal results. In the more difficult Survival Prediction, SP-LNPL still has an improvement on the MIL methods. In the future, we will continue to explore the SP-LNPL considering the cross-domain problem in data cleaning and spatial information in Top-K key instances selection.

Effects of Super Patch Size in SP-LNPL. Ablation results are summarized in Tab. 2. We have following observation: **1)** In the Camelyon16 dataset, different-sized super patches have a promotion for the Tumor Diagnosis. At the same time, we find that a smaller or larger super patch will affect the data-cleaning ability of SP-LNPL. Since the Camelyon16 dataset has a smaller tumor area, a larger super patch will lead to the erroneous deletion of some true positive areas. Besides, the smaller super patch cannot provide enough true negative instances in a super patch, resulting in a weaker classification of false positives. **2)** In the Survival Prediction of CRC-Surv, bag-level prediction performance is related to the selection accuracy and spatial correlation of Top-K key instances. Selecting a moderate-sized super patch like 50 can achieve a better balance between the two factors and thus reach better results.

Effects of Labeled Data at Different Proportions. Ablation results of Top-K key instances selection and λ in *Instance Distribution Aware Task* (IDA-Task) are summarized in Tab. 3 and Tab. 4, respectively. We have the following observations: **1)** A larger Top-K will inevitably introduce more noise in the Top-K key instances selection. However, limited by the low confidence of weak classifiers, a relatively large Top-K can minimize the loss of key instances. Therefore, for the number of Top-K instances selection, we make a tradeoff that chooses 400 at 0.1%/0.5% Labeled and 200 for Top-K at 1% Labeled. **2)** Similarly, the confidence of positive distribution labels in IDA-Task is determined by the performance of weak classifiers. λ determines the regularization of instance-level supervision for bag-level training. We choose a smaller 0.001 at 0.1%/0.5% Labeled and a larger 0.01 at 1% Labeled. Besides, since Camelyon16 is a relatively simple WSI-level binary classification problem, weak regularization can already achieve satisfactory results, so we also choose 0.001 at 0.5% Labeled.

2. Visualization Analysis

As shown in Fig. 1, we visualize the prediction results on TCGA-COAD [5], where the weak classifier is trained on the NCT-CRC-HE dataset [3]. Constrained by the LPA and the cross-domain problem, as shown in Fig. 1b, the ROI regions predicted by the weak classifier have great deviations. As shown in Fig. 1c, SP-LNPL can alleviate the large number of noisy pseudo-labels to a certain extent. Besides, affected by weak classifiers' extremely low prediction confidence, SP-LNPL, as a data cleaning method, still has limitations.

As shown in Fig. 2 and Fig. 3, we visualize the super patch and false positive patches selected by SP-LNPL in Camelyon16 [1] and CRC-Surv, respectively. From Fig. 2a and Fig. 3a, we find that most of the patches in the super patch have great similarity in morphology, indicating that we can effectively cluster the patches in WSI according to the high-dimensional features of the patch. Furthermore, we analyze the false positive patches selected by SP-LNPL in Fig. 2b and Fig. 3b. Due to the weak generalization of weak classifiers, some similar or blurry corrupted images will be mislabeled as positive labels with high confidence, resulting in false positives during the Top-K key instances selection. Specifically, in Fig. 2b, some similar lymphatic

Architecture	Tumor Diagnosis					Survival Prediction				
	0%	0.5%	0.5% (*)	1%	1% (*)	0%	0.1%	0.1% (*)	1%	1% (*)
AB-MIL	$\underline{0.840}_{.024}$	$0.877_{.030}$	$0.932_{.005}$	$0.873_{.006}$	$0.854_{.044}$	$\mathbf{0.582}_{.069}$	$\underline{0.601}^{\dagger}_{.057}$	$\underline{0.601}^{\dagger}_{.057}$	$0.592_{.068}$	$\underline{0.604}^{\dagger}_{.074}$
DS-MIL	$0.743_{.066}$	$0.821_{.058}$	$0.900_{.009}$	$0.792_{.039}$	$0.810_{.018}$	$0.564_{.068}$	$0.552_{.060}$	$0.544_{.055}$	$0.540_{.076}$	$0.579_{.074}$
Patch-GCN	$\mathbf{0.925}_{.020}$	$\mathbf{0.944}_{.005}$	$\underline{0.968}_{.012}$	$\mathbf{0.957}_{.003}$	$\underline{0.969}_{.009}$	$\underline{0.580}_{.024}$	$0.578_{.022}$	$0.590^{\dagger}_{.030}$	$\underline{0.598}_{.042}$	$0.582_{.041}$
LNL-MIL	/	$\underline{0.902}_{.040}$	$\mathbf{0.971}_{.011}$	$\underline{0.944}_{.007}$	$\mathbf{0.986}_{.007}$	/	$\mathbf{0.625}^{\dagger}_{.040}$	$\mathbf{0.627}^{\dagger}_{.043}$	$\mathbf{0.606}^{\dagger}_{.085}$	$\mathbf{0.621}^{\dagger}_{.074}$

Table 1. **Verify the Effects of SP-LNPL Method for MIL Tasks.** We select four representative MIL algorithms: bypass attention based AB-MIL [7], non-local attention based DS-MIL [4], GNN based Patch-GCN [2], and Transformer based LNL-MIL (Ours). **0.1%/0.5%, 1%.** Weak classifier is trained based on LPA, and used to select Top-K key instances. **0.1%(*)/0.5%(*), 1%(*).** Based on the weak classifier, SP-LNPL is adopted. Then, we select Top-K key instances.

Super patch	Tumor Diagnosis		Survival Prediction	
	0.5%	1%	0.1%	1%
w/o	$0.902_{.040}$	$0.944_{.007}$	$0.625^{\dagger}_{.040}$	$\underline{0.606}^{\dagger}_{.085}$
25	$0.924_{.047}$	$0.948_{.019}$	$\mathbf{0.628}^{\dagger}_{.045}$	$0.603^{\dagger}_{.072}$
50	$\mathbf{0.971}_{.011}$	$\mathbf{0.986}_{.007}$	$\underline{0.627}^{\dagger}_{.039}$	$\mathbf{0.621}^{\dagger}_{.074}$
100	$\underline{0.942}_{.010}$	$\underline{0.958}_{.014}$	$0.596^{\dagger}_{.039}$	$0.605^{\dagger}_{.064}$

Table 2. **Effects of the Super Patch Size in SP-LNPL.** We test the effect of different sized super patches in SP-LNPL on the Tumor Diagnosis and Survival Prediction.

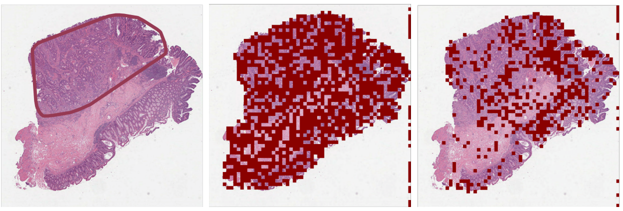
Top-K	Tumor Diagnosis		Survival Prediction	
	0.5%	1%	0.1%	1%
100	$0.968_{.005}$	$0.965_{.008}$	$0.571_{.022}$	$0.602^{\dagger}_{.056}$
200	$\underline{0.969}_{.006}$	$\mathbf{0.986}_{.007}$	$\underline{0.603}^{\dagger}_{.057}$	$\mathbf{0.621}^{\dagger}_{.074}$
400	$\mathbf{0.971}_{.011}$	$\underline{0.906}_{.136}$	$\mathbf{0.627}^{\dagger}_{.043}$	$\underline{0.612}^{\dagger}_{.041}$

Table 3. **Effects of Different Top-K Key Instances Selection.** The confidence of the pseudo-labels predicted by the weak classifier under different proportions of labeled data is distinct. We explore the impact of Top-K selection on two downstream tasks under different proportions of labeled data.

λ	Tumor Diagnosis		Survival Prediction	
	0.5%	1%	0.1%	1%
w/o	$\underline{0.964}_{.006}$	$0.983_{.008}$	$0.603^{\dagger}_{.027}$	$0.614^{\dagger}_{.074}$
0.001	$\mathbf{0.971}_{.011}$	$\mathbf{0.986}_{.007}$	$\mathbf{0.627}^{\dagger}_{.043}$	$\underline{0.616}_{.077}$
0.010	$0.942_{.026}$	$0.982_{.009}$	$\mathbf{0.627}^{\dagger}_{.039}$	$\mathbf{0.621}^{\dagger}_{.074}$

Table 4. **Effects of λ in IDA-Task.** We discuss the selection of λ parameters under different proportions of labeled data.

regions, blurred damaged images, debris regions, etc., will have errors. In Fig. 3b, some similar smooth muscle tissue and cluttered debris areas will have errors.



(a) TCGA-AD-6890.

(b) Weak classifier.

(c) Ours, $t_{\text{ROI}} = 0.5$.

Figure 1. **The Visualization of Weak Classifier Predictions (CRC). (a).** Dark red represents the pathologist-annotated tumor area. We adopt the TCGA cancer coarse annotation from Loeffler *et al.* [6] **(b).** Dark red represents the tumor region predicted by the weak classifier after only FSL. **(c).** Dark red represents the tumor region predicted by the weak classifier after FSL and data cleaning by the SP-LNPL.

References

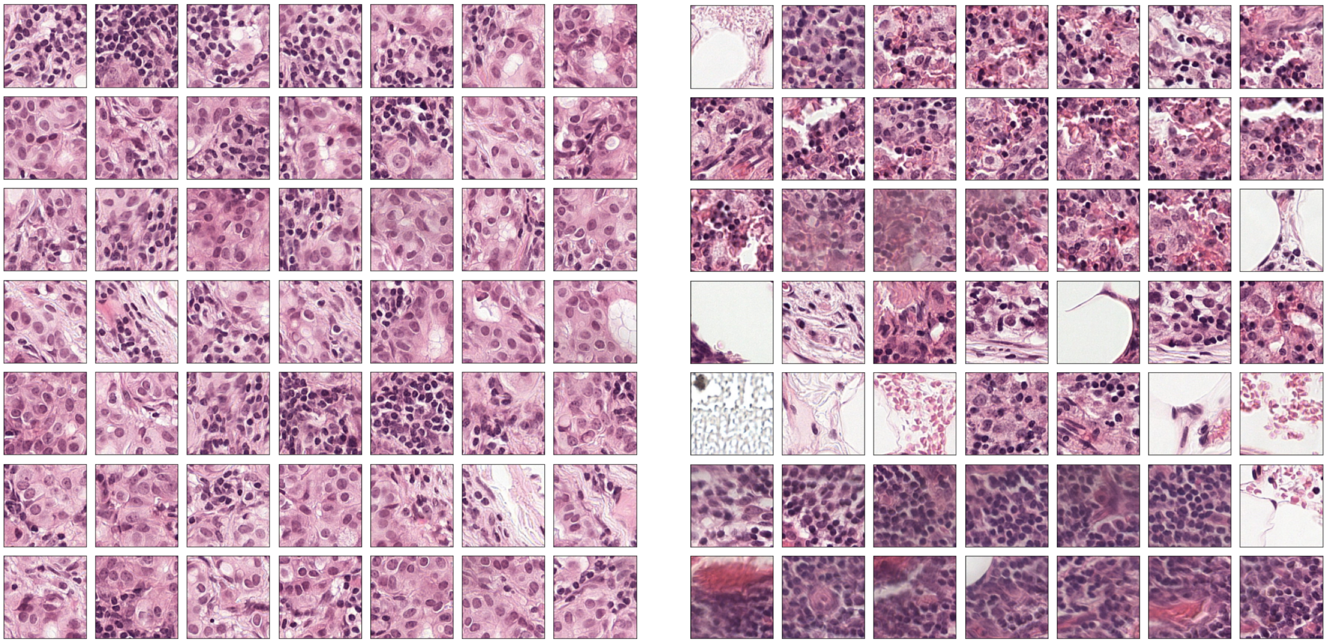
[1] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, Meyke Hermsen, Quirine F Manson, Maschenka Balkenhol, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22):2199–2210, 2017. 1

[2] 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

[3] Jakob Nikolas Kather, Niels Halama, and Alexander Marx. 100,000 histological images of human colorectal cancer and healthy tissue, Apr. 2018. 1

[4] Bin Li, Yin Li, and Kevin W Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 2

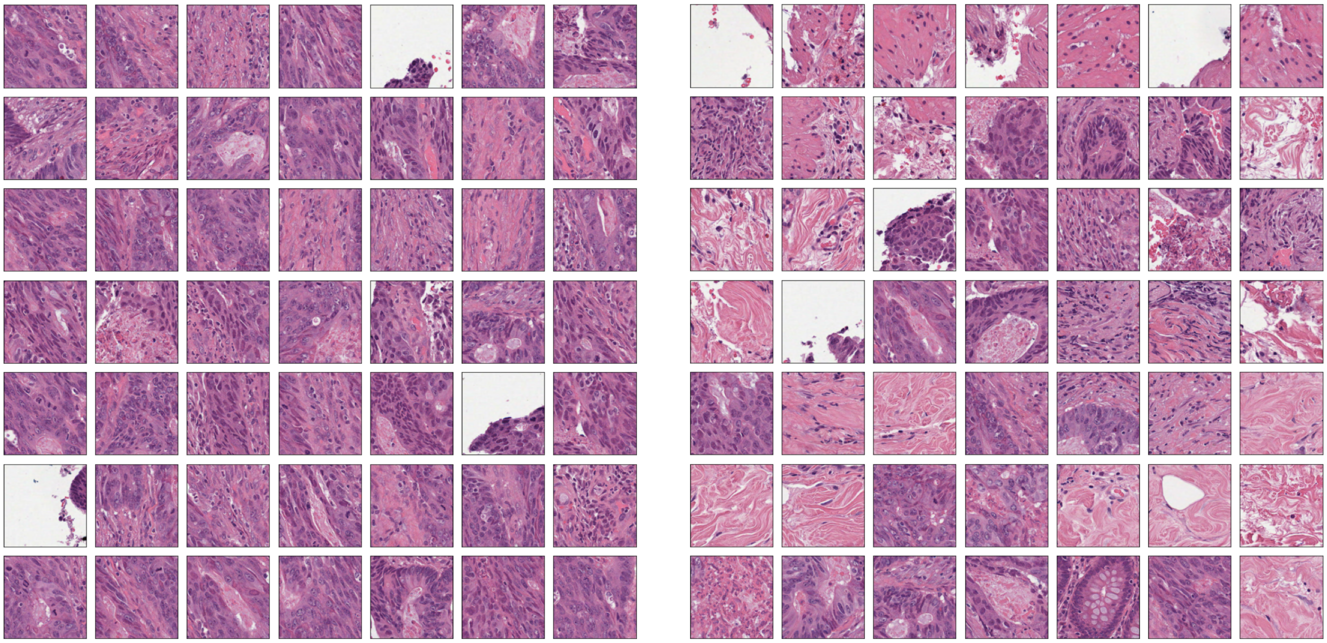
[5] Jianfang Liu, Tara Lichtenberg, Katherine A Hoadley, Laila M Poisson, Alexander J Lazar, Andrew D Cherniack, Albert J Kovatich, Christopher C Benz, Douglas A Levine, Adrian V Lee, et al. An integrated tcga pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell*, 173(2):400–416, 2018. 1



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

(b) Visualization of false positive patches in Top-K key instances.

Figure 2. **The Visualization of Super Patch and False Positive Patches in Top-K Key Instances (Camelyon16). (a).** We visualize the super patch obtained by the KNN search method in the feature space. **(b).** We analyze the false positive pseudo-labels of weak classifiers, and visualize the false positive patches in Top-K key instances.



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

(b) Visualization of false positive patches in Top-K key instances.

Figure 3. **The Visualization of Super Patch and False Positive Patches in Top-K Key Instances (CRC). (a).** We visualize the super patch obtained by the KNN search method in the feature space. **(b).** We analyze the false positive pseudo-labels of weak classifiers, and visualize the false positive patches in Top-K key instances.

[6] Chiara Loeffler and Jakob Nikolas Kather. Manual tumor annotations in tcga, Aug. 2021. 2

[7] Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient

and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5(6):555–570, 2021. 2