西湖大學 | 工学院
WESTLAKE UNIVERSITY | SCHOOL OF ENGINEERING

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Task-Specific Fine-Tuning via Variational Information Bottleneck for Weakly-Supervised Pathology Whole Slide Image Classification

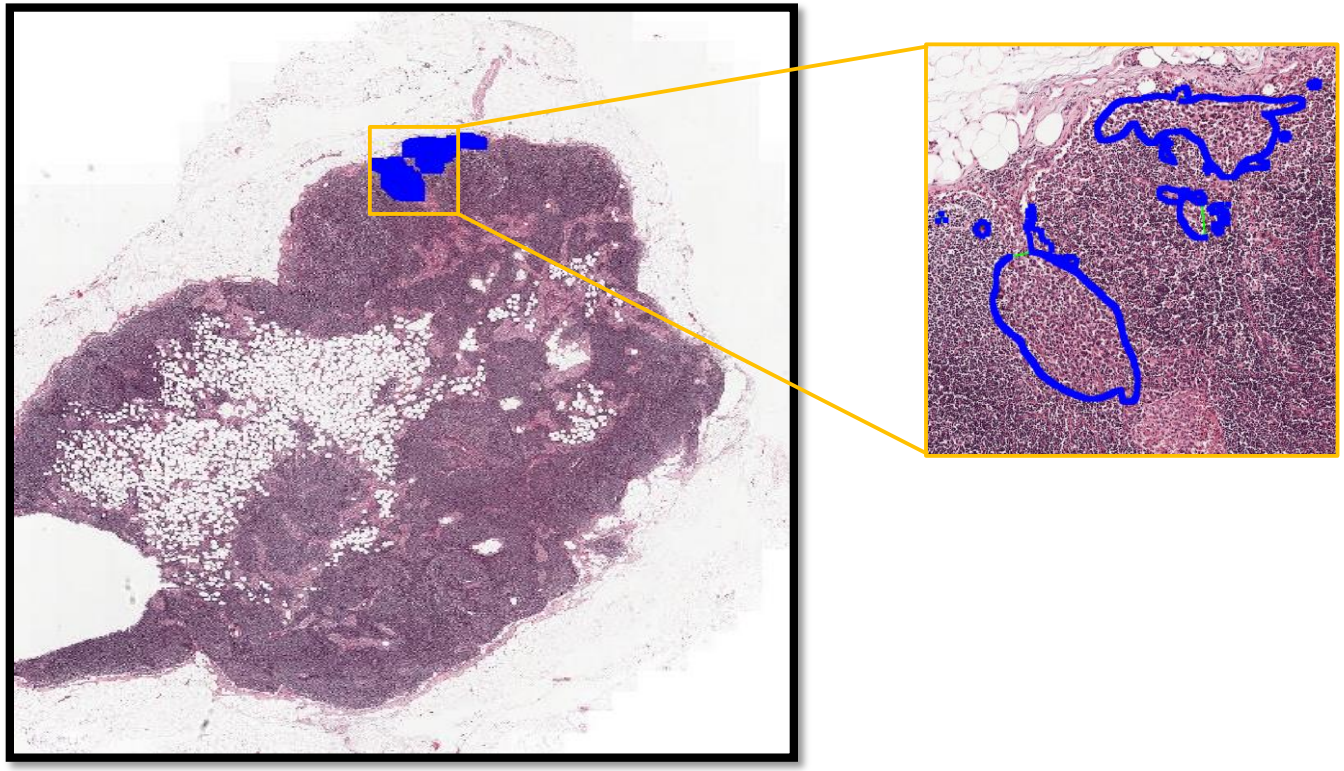
Honglin Li, Chenglu Zhu, Yunlong Zhang, Yuxuan Sun,
Zhongyi Shui, Wenwei Kuang, Sunyi Zheng, Lin Yang

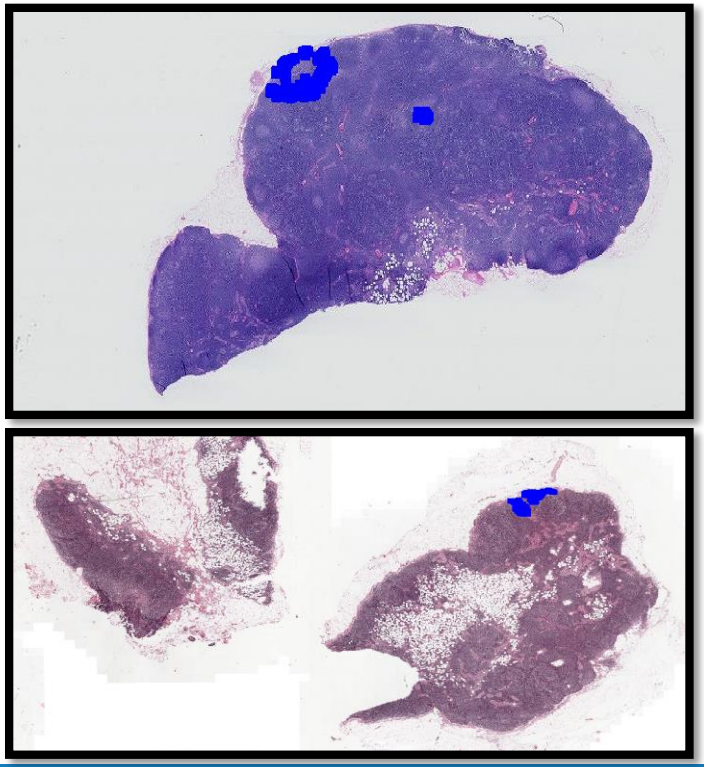
Contents

- Pathology Whole Slide Image Classification
- Related Work
- The Proposed Method
- Experiments

Pathology Whole Slide Image (WSI)

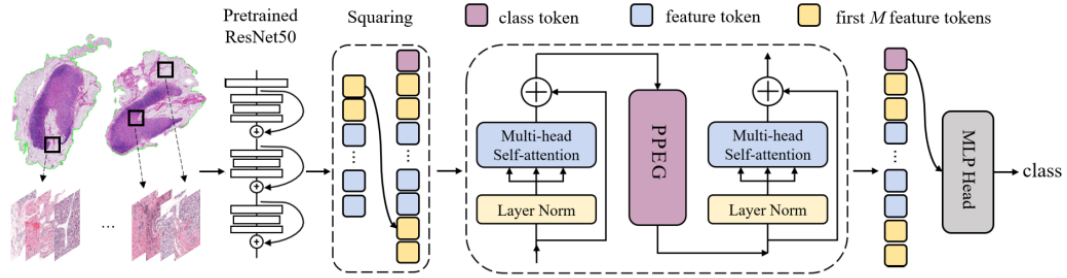
- Lack of patch level annotation (generally slide level diagnosis label will be given)
- Large scale of resolution pixels (hindering parallel training)



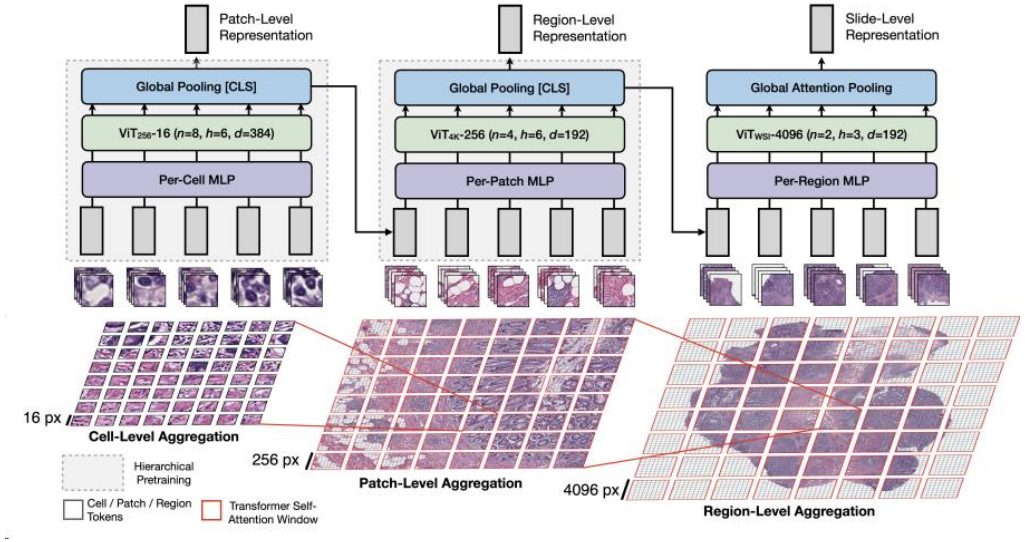


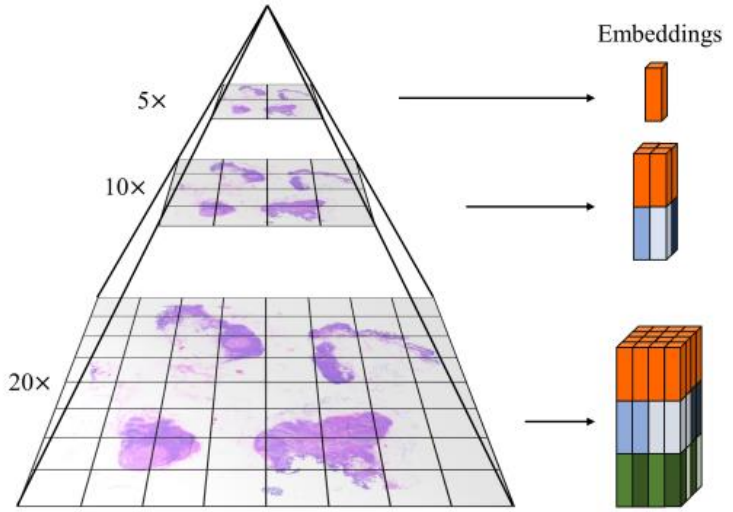
Related Works

- Focus on WSI architecture: CLAM [1], TransMIL [2] …



- Utilize Self-supervised Learning, and multi-scaling: DS-MIL [3], HIPT [4]…





[1] Ming Y Lu, et, al. Data-efficient and weakly supervised computational pathology on whole slide images. Nature Biomedical Engineering 2021.
[2] Zhuchen Shao, et, al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. NeuIPS 2021.
[3] Bin Li, et, al. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. CVPR 2021.
[4] Richard J. Chen and et al. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. CVPR 2022.

Figures in this slide are collected from above papers.

The Proposed Method

- Distilling WSI into simplified bag

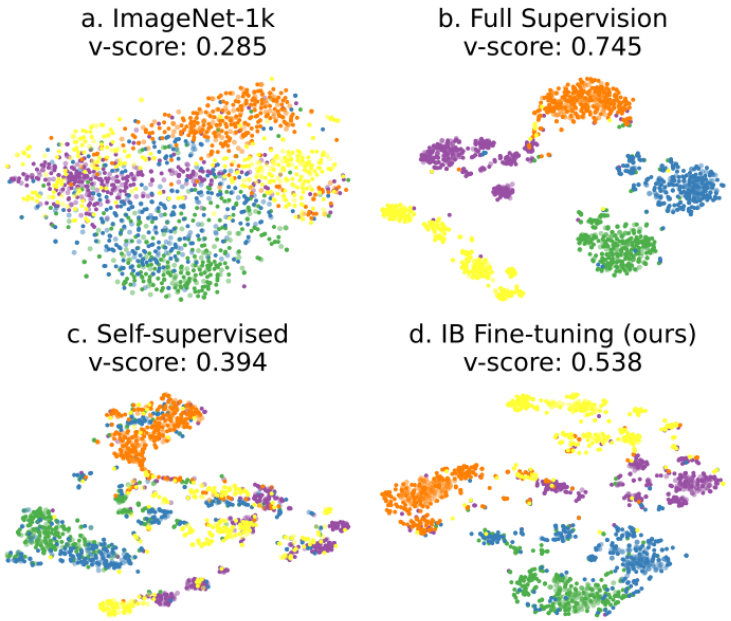
Variational information bottleneck is a useful attributing tool to find the minimal sufficient statistical of WSI.

- Fine-tuning pretrained backbone (after ImageNet transfer or SSL)

utilizes less than a 1% fraction of full patch annotation

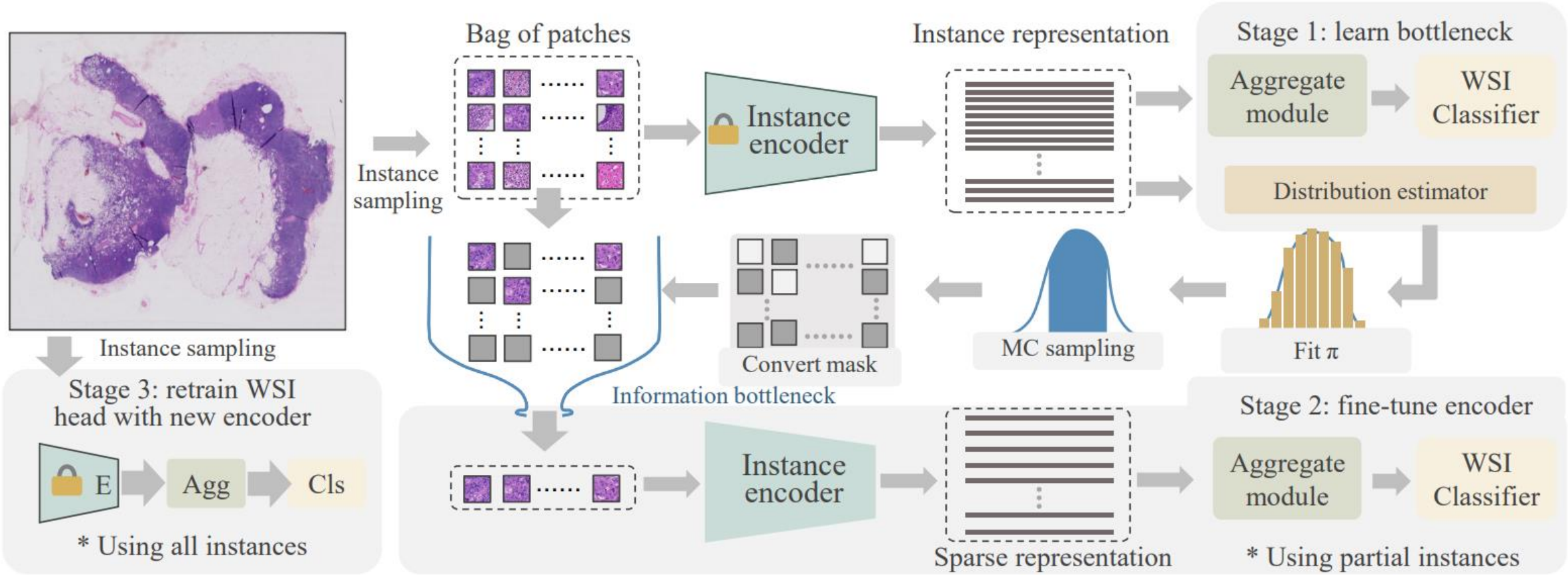
- Versatile training-time augmentations

e.g. Color jitter, rotation are the most common simulation to the variants for digital pathology slides.



The Proposed Method

- Overview of framework



The Proposed Method

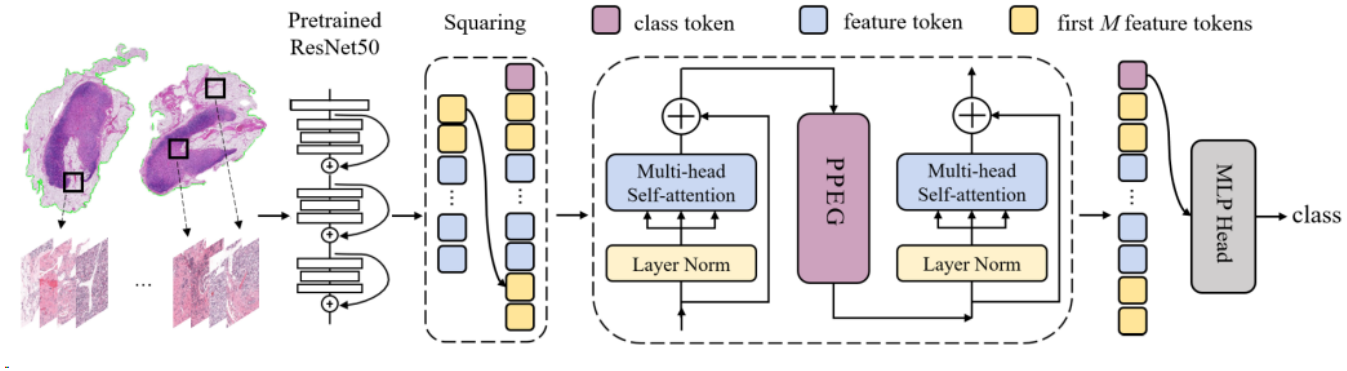
- WSI-MIL definition (original methods)

$$X = \{x_1, ..., x_N\} \qquad \hat{Y} = \max\{\hat{y}_1, ..., \hat{y}_N\}$$

$$Z = \{z_1, ..., z_N\} \qquad z_i = h(x_i; \theta_1)$$

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

$$f(X; \theta) = g\{h(X; \theta_1); \theta_2\}$$



[1] Zhuchen Shao, et, al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. NeuIPS 2021.
Figures in this slide are collected from above paper.

The Proposed Method

- Information Bottleneck for MIL sparsity

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

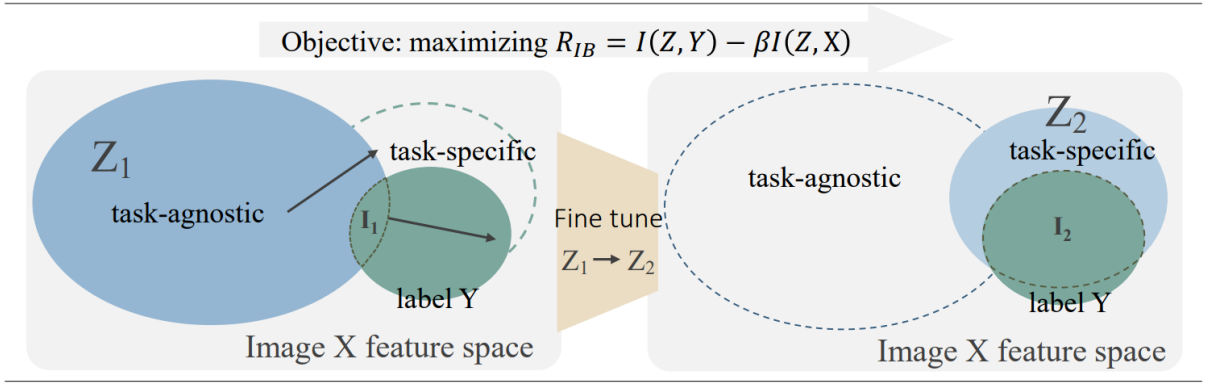
$$J_{IB} = \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}_{z\sim p_\theta(z|x_n)}[-\log q_\phi(y_n|z)] + \beta KL[p_\theta(z|x_n), r(z)],$$

$$z = m \odot x,$$

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

$$P_{set} = \{p(m_1|x_1), ..., p(m_N|x_N)\},$$

$$\hat{Y} = \max\{P_{set}\} = \max\{P_{subset}\},$$

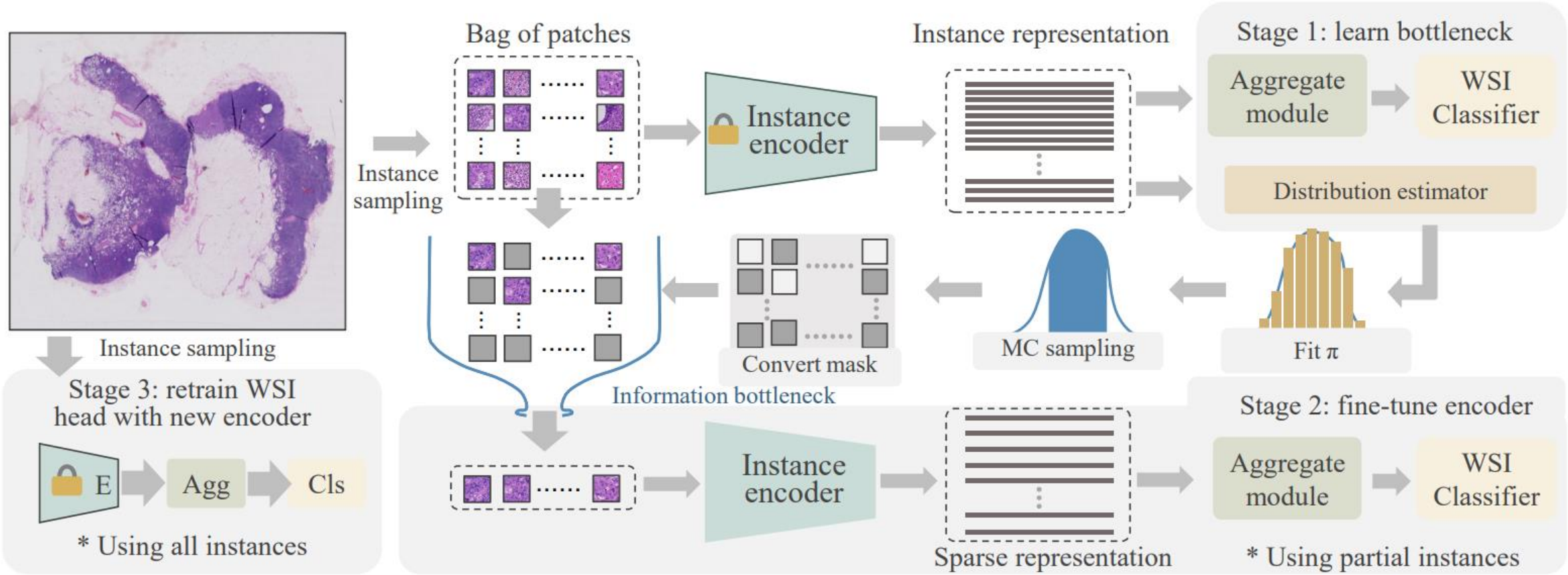


$$loss = \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}_{z\sim p_\theta(z|x_n)}[-\log q_\phi(y_n|z)] + \beta KL[p_\theta(m|x_n), r(m)],$$

[1] Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. ICLR 2017

The Proposed Method

- stages



Experiments

- Datasets

Camelyon-16 [1]: public, metastasis detection in breast cancer

TCGA-BRCA [2]: public, breast invasive carcinoma cohort

LBP-CECA : our private data, cytology for cervical cancer

Camelyon-16-C [3]: generated with random synthetic domain shift

Camelyon-17 [4]: public, metastasis detection in breast cancer

- Runtime, implementations

Algorithm 1: PyTorch-style pseudocode for WSI task-specific IB sparsity learning

```
# Learn sparsity of WSI with fixed
 backbone
for (X,y) in data_loader:
   with torch.no_grad():
      model.eval()
      Z_0 = model(X)
      # X = x_1,x_2,...,x_n
      # Z = z_1,z_2,...,z_n
   model.train()
   # IB is a sequential FCs
   M = IB(Z_0)
   logits = torch.sigmoid(M)
   p_z = Bernoulli(logits)
   Z_mask = p_z.sample()
   r_z = Bernoulli(π)
   # reparameterization trick for
    Bernoulli samples
   Z_1 = Z_0·(M+Z_mask)/2
   Y = model_wsi(Z_1)
   loss1 = CrossEntropyLoss(Y,y)
   loss2 = KL_divergence(p_z, r_z)
   loss = loss1+βloss2
   optimizer.zero_grad()
   loss.backward()
   optimizer.step()
```


[1] Babak Ehteshami Bejnordi, et, al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. Jama 2017.

[2] Nicholas A Petrick, et, al. Spie-aapm-nci breastpathq challenge: an image analysis challenge for quantitative tumor cellularity assessment in breast cancer histology images following neoadjuvant treatment. Journal of Medical Imaging 2021.

[3] Yunlong Zhang, et, al. Benchmarking the robustness of deep neural networks to common corruptions in digital pathology. MICCAI 2022.

[4] Geert Litjens, et al. 1399 h&e-stained sentinel lymph node sections of breast cancer patients: the camelyon dataset. GigaScience 2018.

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

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

Experiments

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

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

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

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

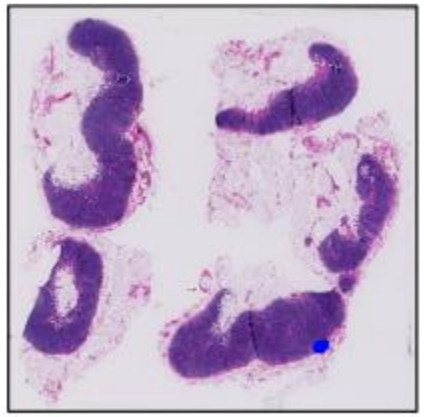
Experiments

- Ablations and visualizations

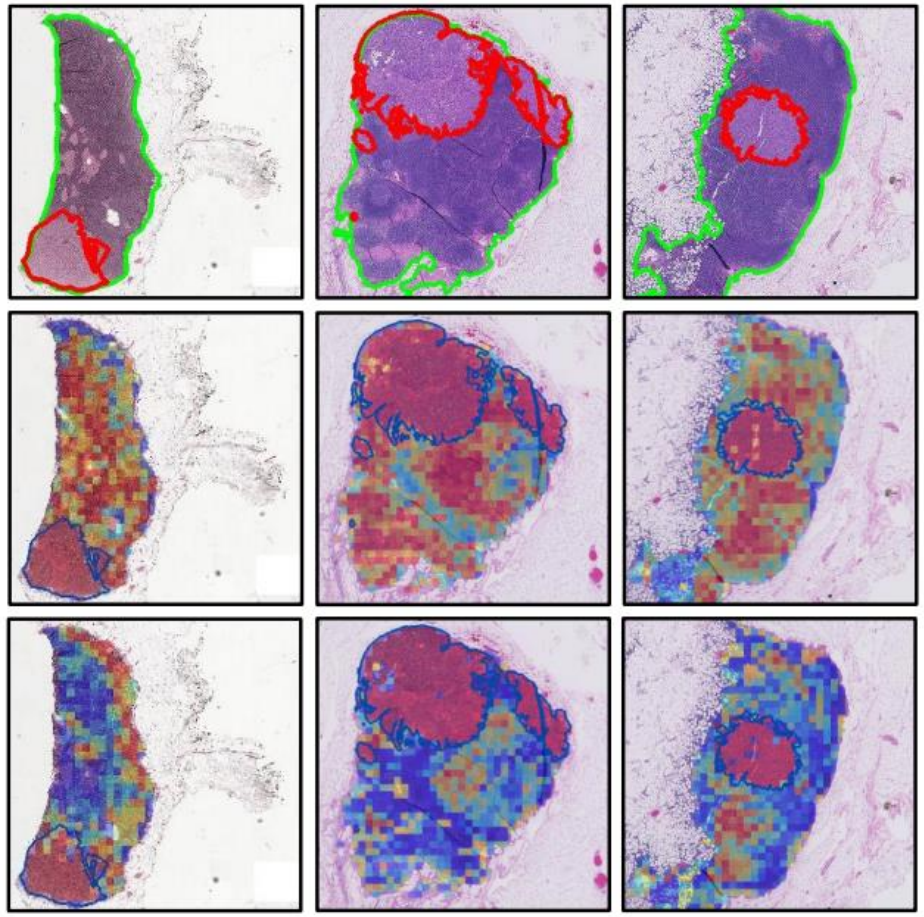
fine-tuning learning rate

LR	F1	AUC
1e-3	N/A	N/A
5e-4	N/A	N/A
1e-4	0.682	0.744
5e-5	0.713	0.741
1e-5	**0.899**	**0.944**
5e-6	0.876	0.908
1e-6	0.806	0.804

3-stages results.



Method	AUC
CLAM-SB	0.875
stage-1	0.865
stage-2	0.944
stage-3	0.956
stage-2 random	0.731



Experiments

- Ablations

Top-k and beta selection during VIB training.

Top-K	F1	AUC
128	0.840±0.011	0.870±0.010
256	0.843±0.009	0.870±0.010
512	0.843±0.005	0.866±0.011
1024	0.845±0.007	0.864±0.011
2048	**0.846±0.004**	**0.875±0.010**
all	0.839±0.018	0.875±0.028

β	F1	AUC
Upper bound	0.839±0.018	0.875±0.028
1e-3	0.835±0.008	0.860±0.012
1e-2	0.833±0.006	0.860±0.028
1e-1	**0.849±0.010**	**0.865±0.014**
1	0.839±0.015	0.852±0.018
10	0.838±0.016	0.862±0.020
100	0.828±0.010	0.853±0.007

西湖大學 | 工学院
WESTLAKE UNIVERSITY | SCHOOL OF ENGINEERING

THANKS