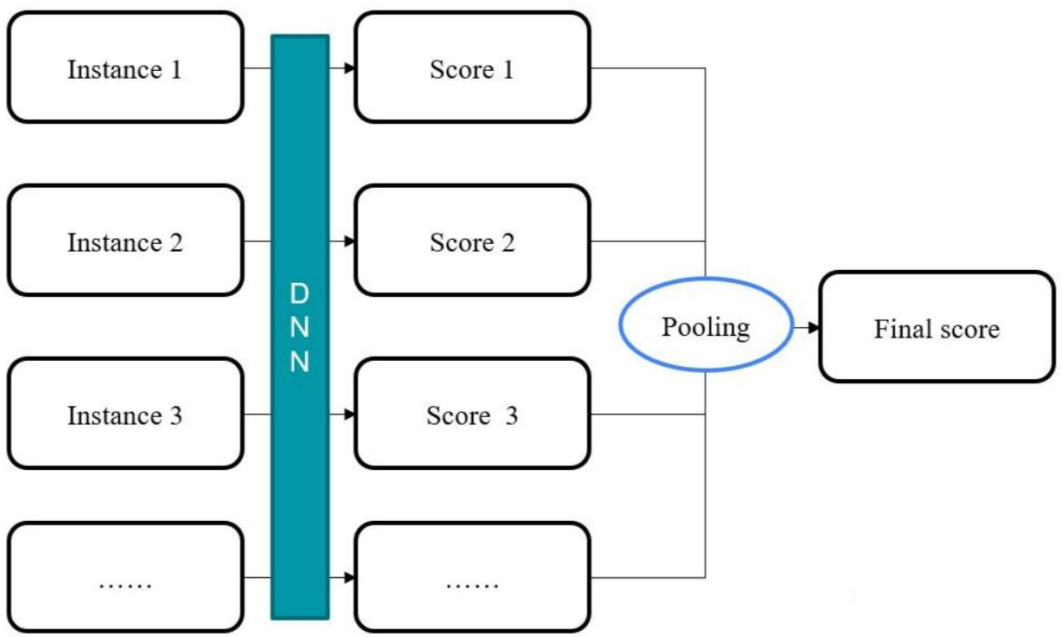
Fast and Accurate Gigapixel Pathological Image Classification with Hierarchical Distillation Multi-Instance Learning

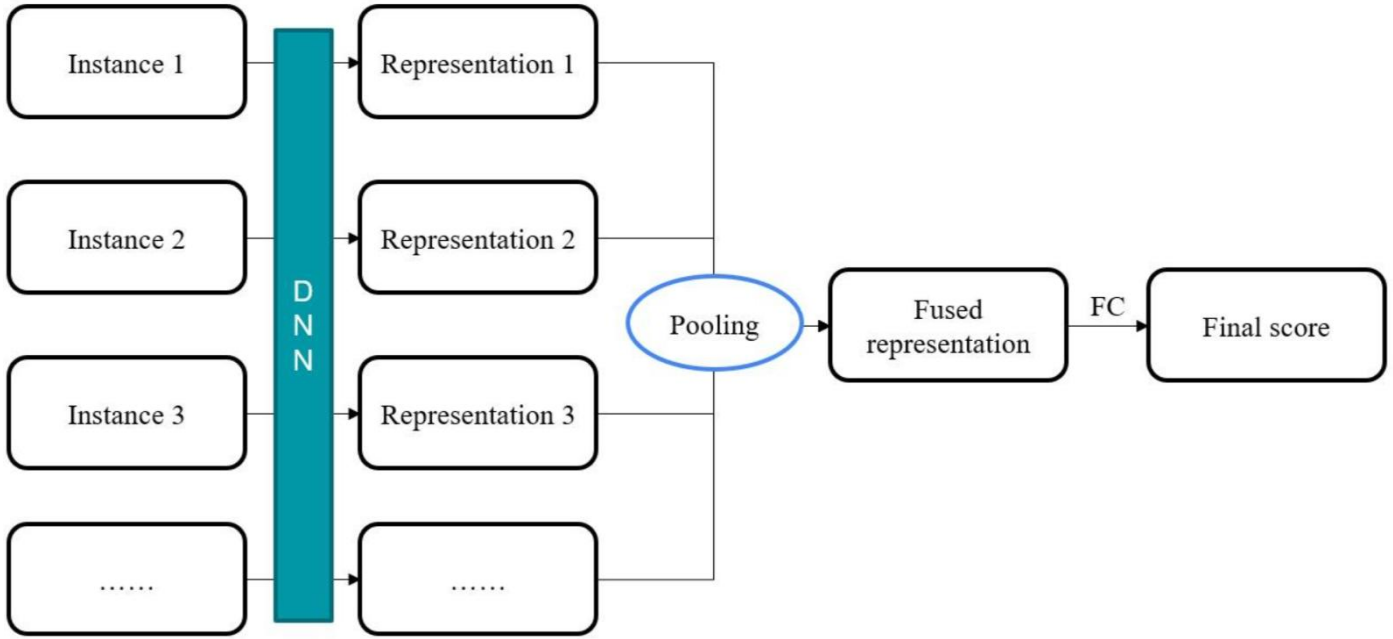
Jiuyang Dong[1], Junjun Jiang[1]*, Kui Jiang[1], Jiahan Li[1],
Yongbing Zhang[2]*
[1]Harbin Institute of Technology, [2]Harbin Institute of Technology, Shenzhen
{jiuyang.dong, jiahan.li}@stu.hit.edu.cn,
{jiangjunjun, jiangkui, ybzhang08}@hit.edu.cn

accepted by CVPR 2025

Multi-instance Learning (MIL)



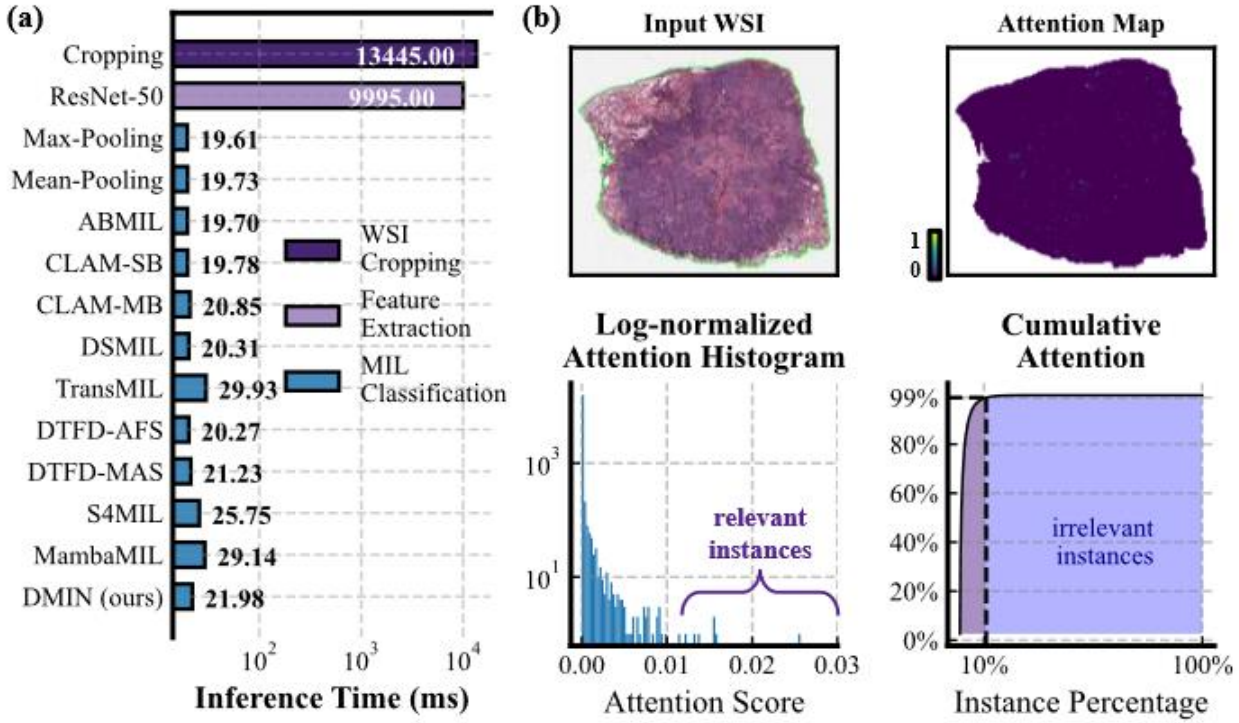
Instance-based approach



Embedding-based approach

Motivation

Multi-instance Learning (MIL)



Time-consuming data pre-processing:
After comparing the time required for data pre-processing (WSI cropping, feature extraction) and MIL network classification, it is clear that data pre-processing is the main speed bottleneck.

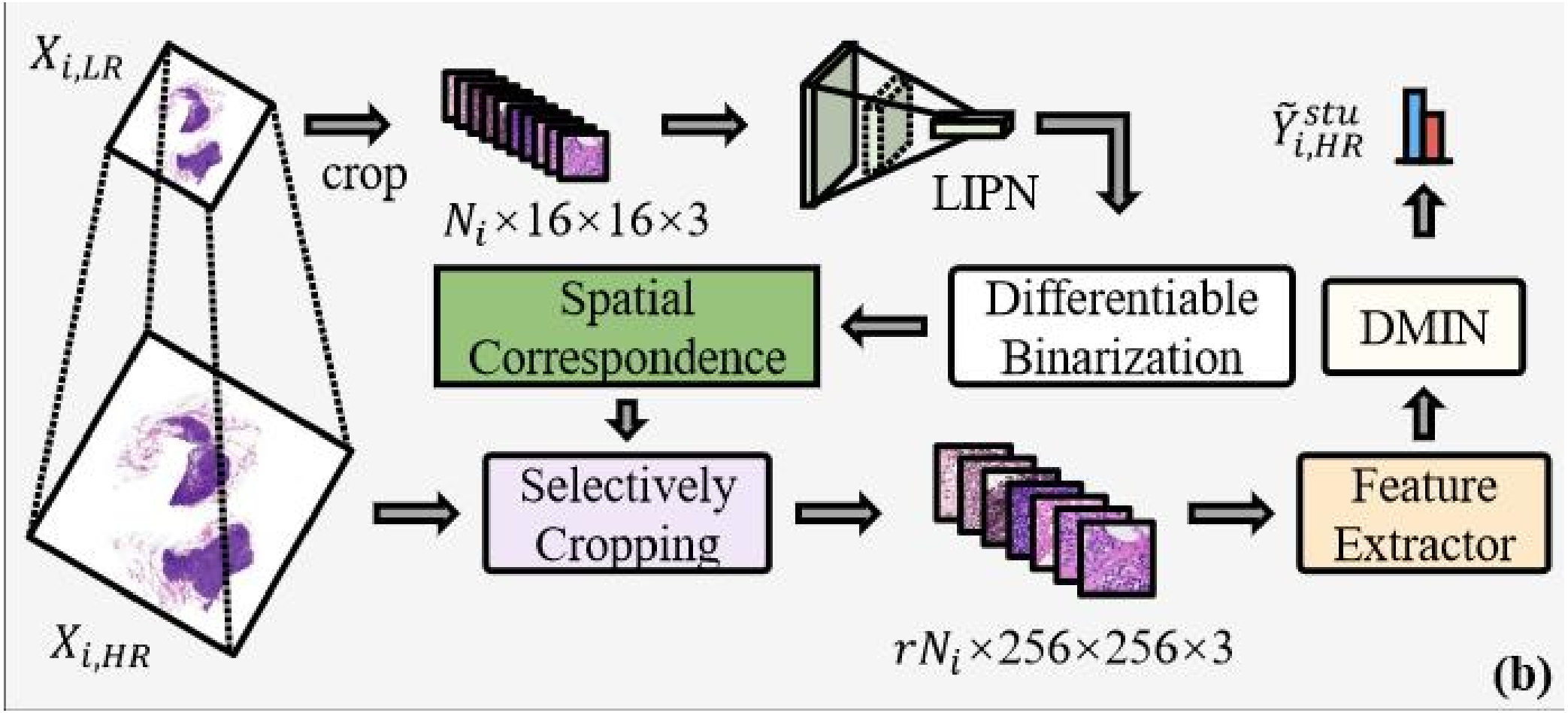
Redundant irrelevant patches:
For example, in a randomly selected WSI, numerous instances have extremely low attention scores, indicating their minimal contribution, if any, to the bag-level classification.

Based on the above analysis, a straightforward idea to reduce the inference time is discarding irrelevant instances based on attention scores.

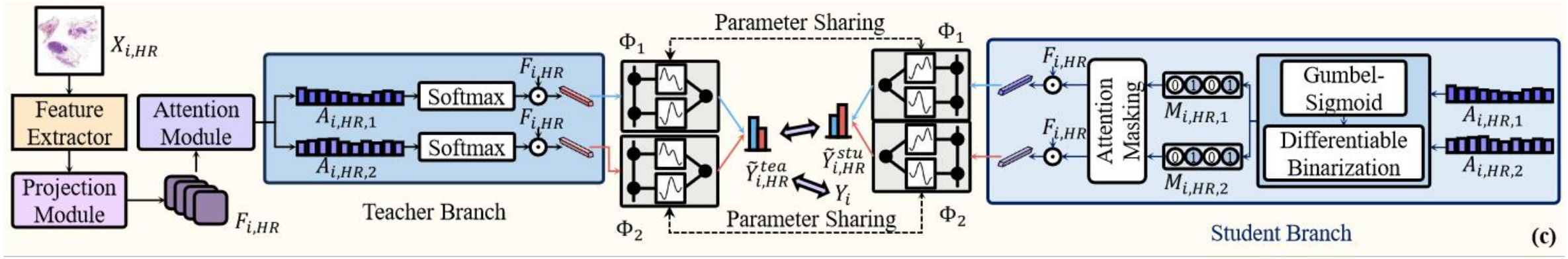
Related work

SMT employs cascading vision transformer (ViT) blocks to gradually search for “suspicious” areas and ultimately uses only a small area of the entire WSI for classification. The classification performance of SMT heavily relies on accurately identifying potential tumor areas. However, the pathological information provided by the low-resolution thumbnails, used as the initial input of SMT, is insufficient, which can easily lead to inappropriate regions of interest being focused. Consequently, the accumulation of errors results in inferior classification performance of SMT when compared to other non-accelerated MIL methods.

Method


$X_{i,LR}$
crop
$N_i \times 16 \times 16 \times 3$
LIPN
$\tilde{Y}^{stu}_{i,HR}$
Spatial Correspondence
Differentiable Binarization
DMIN
Selectively Cropping
Feature Extractor
$X_{i,HR}$
$rN_i \times 256 \times 256 \times 3$
(b)

Method



Projection and Attention Module

feature extractor: $X_{i,HR} \rightarrow I_{i,HR}$

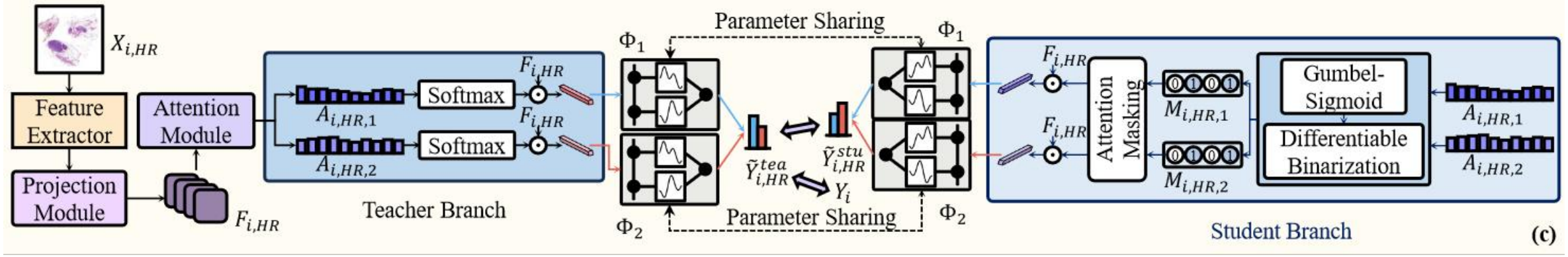
projection: $I_{i,HR} \rightarrow F_{i,HR}$

attention scores: $A_{i,HR} = [\phi(F_{i,HR}V) \odot \sigma(F_{i,HR}U)]W,$

Teacher Branch

bag-level representation: $E^{tea}_{i,HR,c} = \varphi(A_{i,HR,c})^{\top} \otimes F_{i,HR,c}, c \in \{1, 2\}.$

Method



Student Branch and Self-Distillation

the student branch is designed to compute bag-level representations using only a subset of instances with larger attention scores

we incorporate the Gumbel Noise to "sigmoid" the un-normalized attention matrices:

$$\hat{A}_{i,HR,c} = \sigma\left(\frac{A_{i,HR,c} + G_{1,c} - G_{2,c}}{\tau}\right), c \in \{1, 2\}.$$

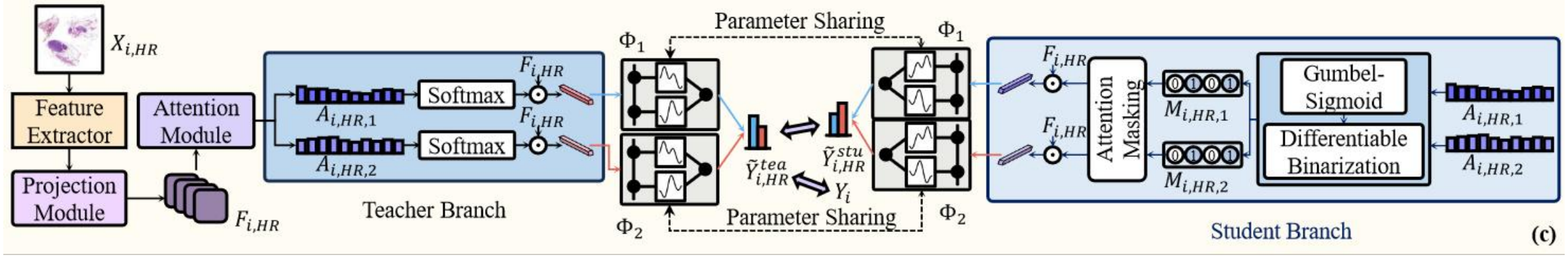
differentiable binarization:

$$M^j_{i,HR,c} = B(\hat{A}^j_{i,HR,c}, \gamma) - D(\hat{A}^j_{i,HR,c}) + \hat{A}^j_{i,HR,c},$$

bag-level representations:

$$E^{stu}_{i,HR,c} = \sum_{j=1}^{N_i} \frac{exp(A^j_{i,HR,c}) M^j_{i,HR,c}}{\sum_{s=1}^{N_i} exp(A^s_{i,HR,c}) M^s_{i,HR,c}} F^j_{i,HR,c},$$

Method



CKA Classifier

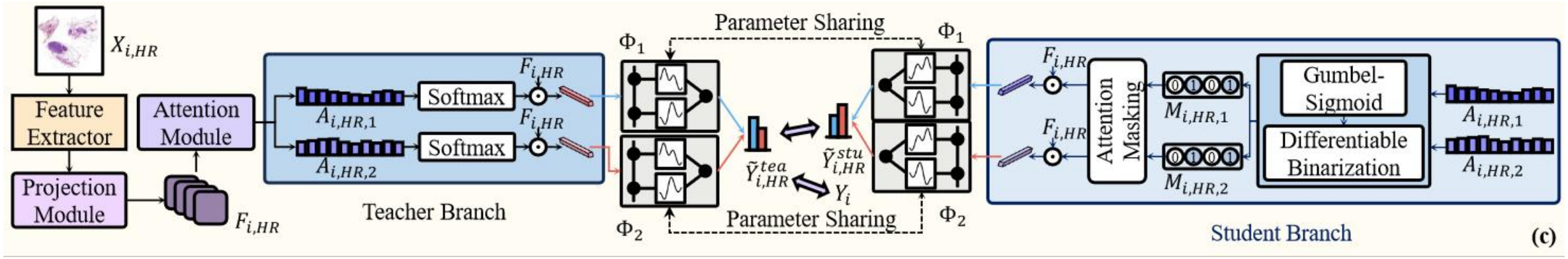
we employ the iterative form of K-order Chebyshev polynomials to represents the basis functions TK(x): $T_K(x) = 2xT_{K-1}(x) - T_{K-2}(x), K \geq 2.$

Here, $x \in \mathbb{R}^{1\times Q}$ represents a bag-level representation, $T_0(x) = \vec{1}$ and $T_1(x) = x.$

prediction: $$\Phi(x)[o] = \sum_{k=0}^{K}\sum_{q=1}^{Q} T_k(x)[q] * \Omega[q, o, k],$$

$$\begin{cases} \widetilde{Y}^{tea}_{i,HR} = [\Phi_1(\phi(E^{tea}_{i,HR,1})) \oplus \Phi_2(\phi(E^{tea}_{i,HR,2}))]. \\ \widetilde{Y}^{stu}_{i,HR} = [\Phi_1(\phi(E^{stu}_{i,HR,1})) \oplus \Phi_2(\phi(E^{stu}_{i,HR,2}))]. \end{cases}$$

Method

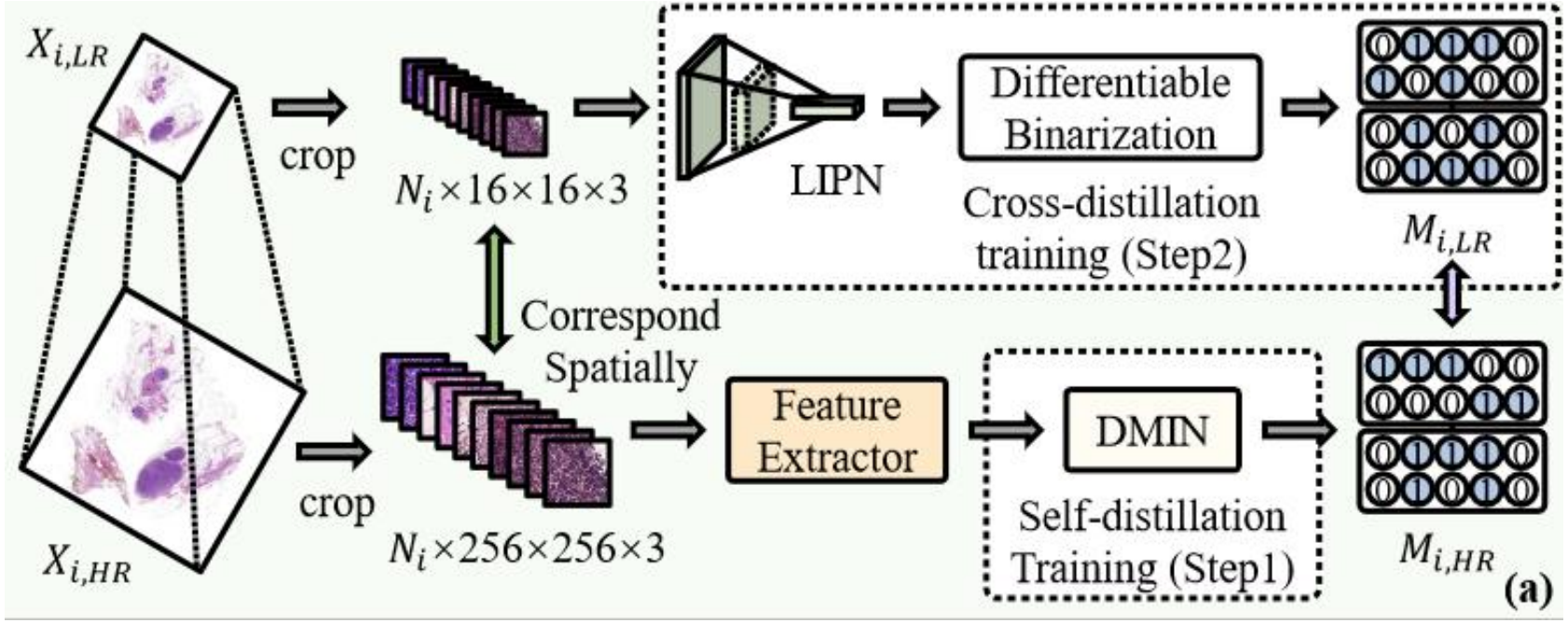


Hybrid Loss Function

$$L^{tea}_{cls} = CE(\tilde{Y}^{tea}_{i,HR}, Y_i), \qquad \begin{cases} L^{stu}_{dis,1} = L_2(E^{stu}_{i,HR}, E^{tea}_{i,HR}), \\ L^{stu}_{dis,2} = L_{KL}(\tilde{Y}^{stu}_{i,HR}, \tilde{Y}^{tea}_{i,HR}). \end{cases} \qquad L^{stu}_{rate} = L_2(\tilde{r}_{i,HR}, r).$$

$$L_{DMIN} = \alpha_1 L^{tea}_{cls} + \alpha_2 L^{tea}_{clu} + \alpha_3 L^{stu}_{dis,1} + \alpha_4 L^{stu}_{dis,1} + \alpha_5 L^{stu}_{rate}.$$

Method



Specifically, the N_i 16×16 patches obtained from $X_{i,LR}$ are directly fed into LIPN, generating dual-branch prediction matrices $P_{i,LR,c}$, $c \in \{1, 2\}$

$$M^j_{i,LR,c} = B(P^j_{i,LR,c}, \gamma) - D(P^j_{i,LR,c}) + P^j_{i,LR,c}.$$

$$L_{LIPN} = \beta_1 \sum_{c=1}^{2} \frac{L_1(M_{i,LR,,c}, M_{i,HR,c})}{2} + \beta_2 L_2(\tilde{r}_{i,LR}, r).$$

Experiment

Comparative Methods	Camelyon16			TCGA-NSCLC			TCGA-BRCA		
	AUC↑	ACC↑	Time(s)↓	AUC↑	ACC↑	Time(s)↓	AUC↑	ACC↑	Time(s)↓
Max-Pooling	$83.26_{1.54}$	$82.41_{0.73}$	23.46	$94.66_{2.33}$	$86.40_{3.73}$	57.16	$88.03_{7.76}$	$86.05_{3.88}$	36.49
Mean-Pooling	$61.80_{2.15}$	$70.54_{1.41}$	23.46	$92.82_{3.54}$	$84.93_{4.78}$	57.16	$88.23_{5.67}$	$86.74_{2.44}$	36.49
ABMIL [19]	$84.88_{3.38}$	$82.79_{2.68}$	23.46	$94.92_{2.29}$	$88.03_{3.65}$	57.16	$87.70_{6.15}$	$87.68_{3.51}$	36.49
CLAMSB [34]	$83.49_{4.46}$	$79.61_{4.40}$	23.46	$95.05_{2.72}$	$88.74_{3.39}$	57.16	$88.25_{6.12}$	$87.58_{4.92}$	36.49
CLAMMB [34]	$87.51_{3.23}$	$82.56_{3.11}$	23.46	$95.59_{2.16}$	$88.01_{3.38}$	57.16	$90.22_{5.18}$	$88.27_{3.52}$	36.49
DSMIL [27]	$75.94_{10.81}$	$75.35_{6.12}$	23.46	$92.11_{2.97}$	$83.67_{3.80}$	57.16	$83.33_{7.48}$	$82.59_{3.66}$	36.49
TransMIL [47]	$82.26_{5.67}$	$81.01_{6.85}$	23.47	$94.57_{2.03}$	$88.21_{3.04}$	57.17	$88.33_{5.73}$	$87.55_{3.78}$	36.49
DTFDAFS [61]	$87.40_{3.17}$	$85.12_{2.42}$	23.46	$95.59_{2.08}$	$88.76_{3.89}$	57.16	$87.24_{7.38}$	$86.83_{3.98}$	36.49
DTFDMAS [61]	$87.75_{2.07}$	$85.43_{2.03}$	23.46	$95.02_{2.32}$	$89.02_{3.78}$	57.17	$87.80_{9.65}$	$87.48_{4.13}$	36.49
S4MIL [10]	$86.40_{1.99}$	$80.39_{2.79}$	23.47	$96.19_{1.89}$	$89.69_{2.86}$	57.17	$90.40_{5.73}$	$88.17_{3.88}$	36.49
MambaMIL [57]	$87.06_{6.19}$	$83.26_{2.93}$	23.47	$95.37_{1.70}$	$89.62_{3.13}$	57.16	$89.69_{5.91}$	$87.78_{4.27}$	36.49
HDMIL†	**$93.17_{1.83}$**	**$88.92_{2.51}$**	**23.46**	**$96.47_{2.20}$**	**$89.75_{2.86}$**	**57.16**	**$90.43_{4.86}$**	**$88.68_{3.17}$**	**36.49**
HDMIL	**$90.88_{2.75}$**	**$88.61_{2.04}$**	**16.75**	**$96.35_{2.26}$**	**$89.78_{3.11}$**	**44.71**	**$90.45_{4.42}$**	**$88.27_{2.47}$**	**33.86**

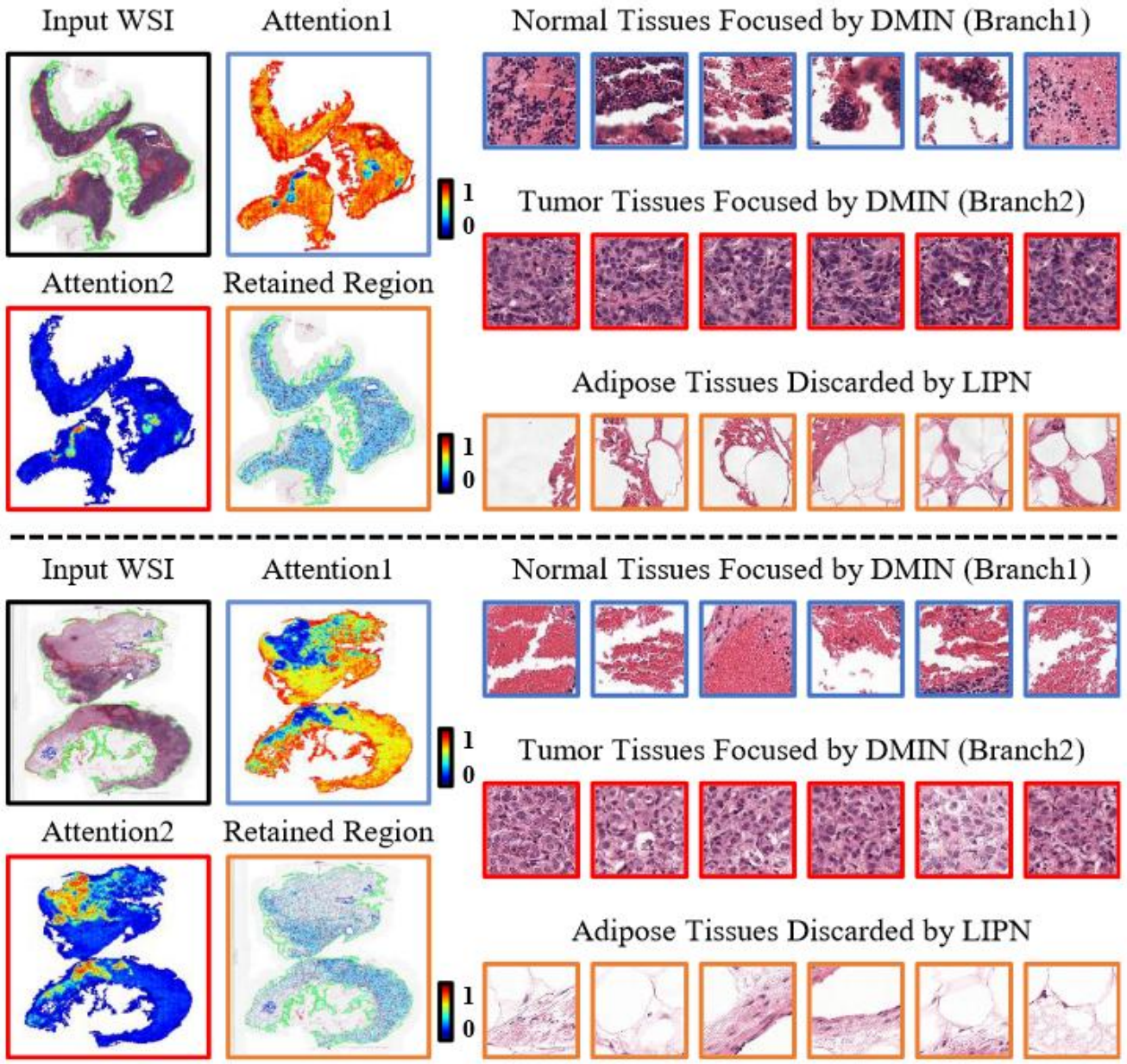
Table 1. Comparison of HDMIL with the state-of-the-art MIL methods on Camelyon16, TCGA-NSCLC, and TCGA-BRCA. The 10-fold **test** AUC and accuracy (ACC) scores are reported in the form of mean_{std}. The best and second best results are indicated in red and blue, respectively. The average processing time per WSI on each test sets are also shown. HDMIL† means using only DMIN for inference.

Experiment

Methods	Dataset	LIPN	Crop	Fea	DMIN	Total
Came16	HDMIL†	-	13.45	10.00	0.02	23.46
	HDMIL	0.01	10.88	5.84	0.02	16.75
	Δ	-	**−19.1%**	**−41.6%**	-	**−28.6%**
NSCLC	HDMIL†	-	47.02	10.12	0.02	57.16
	HDMIL	0.01	37.21	7.48	0.02	44.71
	Δ	-	**−20.9%**	**−26.1%**	-	**−21.8%**
BRCA	HDMIL†	-	27.17	9.30	0.02	36.49
	HDMIL	0.01	25.84	8.00	0.02	33.86
	Δ	-	**−4.90%**	**−14.0%**	-	**−7.2%**

Table 2. Comparison of HDMIL and HDMIL† when splitting the inference time (seconds) into four stages: instance pre-screening (LIPN), WSI cropping ("Crop"), feature extraction ("Fea"), and bag classification (DMIN).

Experiment


Input WSI
Attention1
Normal Tissues Focused by DMIN (Branch1)
1
0
Tumor Tissues Focused by DMIN (Branch2)
Attention2
Retained Region
Adipose Tissues Discarded by LIPN
1
0
Input WSI
Attention1
Normal Tissues Focused by DMIN (Branch1)
1
0
Tumor Tissues Focused by DMIN (Branch2)
Attention2
Retained Region
Adipose Tissues Discarded by LIPN
1
0

Experiment

DMIN		LIPN	Camelyon16		TCGA-NSCLC		TCGA-BRCA		Average	
CKA	SelfDist		AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC
✗	✗	✗	$94.67_{4.51}$	$91.54_{5.38}$	$95.36_{3.51}$	$89.44_{4.51}$	$88.82_{6.41}$	$86.47_{4.21}$	92.95	89.15
✓	✗	✗	$97.15_{3.27}$	$93.85_{4.86}$	$95.19_{3.01}$	$89.67_{3.57}$	$91.22_{5.40}$	**$89.00_{3.62}$**	94.52	90.84
✓	✓	✗	**$97.70_{2.54}$**	**$95.00_{4.81}$**	**$95.58_{3.27}$**	**$90.29_{3.90}$**	**$93.33_{4.58}$**	**$89.83_{2.71}$**	**95.54**	**91.71**
✓	✓	✓	**$97.64_{2.93}$**	**$95.38_{3.97}$**	**$95.88_{3.02}$**	**$90.50_{3.44}$**	**$93.27_{4.87}$**	$88.70_{3.92}$	**95.60**	**91.53**

Table 3. The effect of each component in HDMIL on classification performance. The 10-fold **validation** AUC and ACC scores are reported in the form of mean_{std}. "SelfDist" is the abbreviation for self-distillation.

Thanks