WsiSSM: A Weakly Supervised Subset-Matching Framework for Unified Classification and Segmentation of Histopathology Whole Slide Images

Guangjian Zeng, Siyuan Tao, Chengzhi Zhao, Xiaoqing Li, Na Tang, Wenting Huang, Shenying Fang*
Southern University of Science and Technology

NEURAL INFORMATION PROCESSING SYSTEMS

Code

Paper

Motivation

Weakly supervised whole-slide analysis represents each gigapixel WSI as a variable-length bag of patch embeddings with slide-level supervision. Transformer-based MIL captures inter-instance dependencies but incurs quadratic $O(N^2)$interaction complexity for N patches [1]. Scalable WSI modeling therefore requires sparse attention [2] that accommodates both local tissue morphology and spatially dispersed diagnostic evidence.

- **Gigapixel WSIs**
- **Inter-patch Dependencies**
- **Quadratic Complexity**

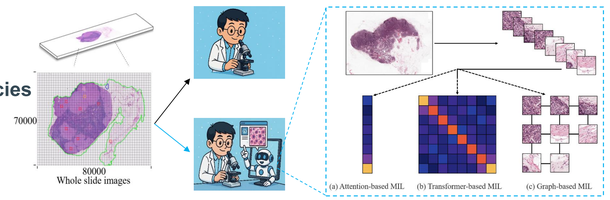


Figure 1. Whole-slide MIL formulations.

Foundation-model patch embeddings retain tumor–benign overlap in the illustrated feature projection (Figure 2). Meanwhile, slide-level classification objectives leave lesion extent underconstrained, allowing spatially incomplete or off-target activations despite accurate bag prediction. These limitations motivate contextual instance modeling and explicit CAM consistency for weakly supervised segmentation.

- **Feature Overlap**
- **Localization Gap**
- **Spurious Activations**

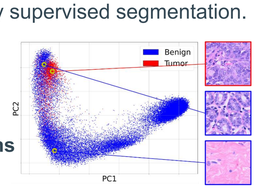


Figure 2. Projected patch-feature distributions.

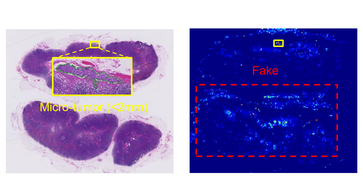


Figure 3. Schematic localization failure modes under slide-level supervision

Method

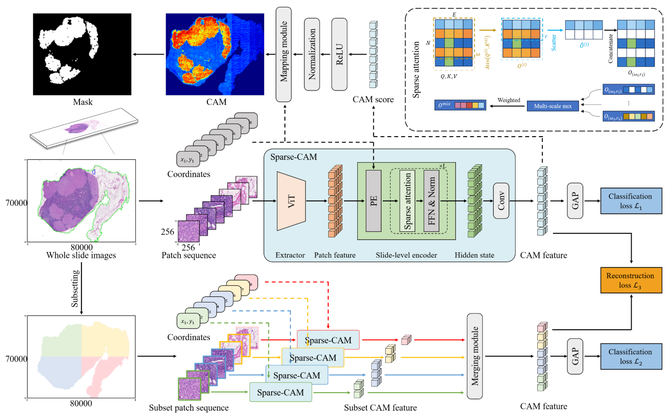


Figure 4. The pipeline of our proposed WsiSSM for WSI.

Sparse-CAM contextualizes frozen foundation-model embeddings through multi-scale sparse self-attention. A shared pointwise classifier and global average pooling jointly generate class-specific CAMs and slide-level predictions. **Subset matching** imposes a consistency constraint between whole-slide CAMs and coordinate-reassembled CAMs from four unequal-length spatial subsets, coupling classification supervision with localization regularization.

Sparse-CAM. Multi-scale sparse attention aggregates complementary patch interactions across window sizes and sampling strides:

$$O^{mix} = \sum_{i=1}^{k} \frac{a_j}{\sum_{j=1}^{k} a_j} O_{(\omega_i,\gamma_i)}, a_j = \sum_{i=1}^{N} exp(W_a O_{(\omega_i,\gamma_i),i} + b), \qquad O^{(i)} = softmax(\frac{Q^{(i)} K^{(i)T}}{\sqrt{E}}) V^{(i)} \quad (1)$$

A shared pointwise classifier couples class-specific activation mapping with slide-level discrimination through global average pooling:

$$\hat{Y} = \sigma(GAP(A)), \qquad A_i = \sum_{k=0}^{K-1} \theta_k F_{i+k} + b \quad (2)$$

Subset matching. Four unequal-length spatial subsets share the same CAM network; coordinate-based reassembly restores patchwise correspondence with the whole-slide prediction:

$$A_s = \mathrm{Merge}_P \left(\{ f_\theta(G_k, P_k) \}_{k=1}^{4} \right), \qquad A = f_\theta(G, P). \quad (3)$$

Dual classification losses supervise both aggregation paths, while \(\ell_1\) CAM consistency regularizes spatial localization without dense annotations.

$$\mathcal{L} = \underbrace{\mathrm{CE}(\hat{y}, y)}_{\mathcal{L}_1} + \underbrace{\mathrm{CE}(\hat{y}_s, y)}_{\mathcal{L}_2} + \alpha \underbrace{\|A - A_s\|_1}_{\mathcal{L}_3}. \quad (3)$$

Mechanism

Classification-only training yields context-dependent CAM intensities between whole-slide and subset inference. Subset matching penalizes activation discrepancies at corresponding patch coordinates, promoting consistent lesion responses across contextual scales. The observed reduction in subset-dependent activation contrast supports CAM consistency as a spatial regularizer under slide-level supervision.

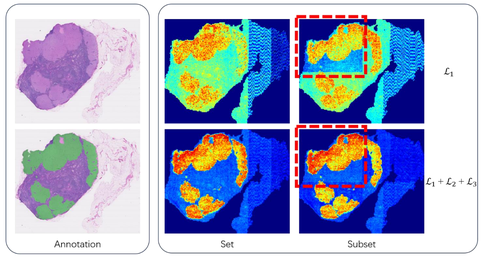


Figure 5. Visualization of WsiSSM's mechanism.

Experiments

Five cohorts comprise 3,666 WSIs: **CAMELYON** (899; breast-cancer lymph-node metastasis), **PANDA** (692; prostate cancer), **NCPDC** (768; primary breast frozen sections), and **two external** LNM cohorts (743/564).

Evaluation separates slide-level discrimination (AUC, sensitivity, specificity) from region-level delineation (Dice, IoU, accuracy). WsiSSM achieves the highest performance across all five cohorts (Table 1).

Table 1. Classification performance [%] of various MIL methods across multiple datasets.

Method	CAMELYON			PANDA			NCPDC			External A			External B		
	Sensitivity	Specificity	AUC	Sensitivity	Specificity	AUC	Sensitivity	Specificity	AUC	Sensitivity	Specificity	AUC	Sensitivity	Specificity	AUC
WOTC	93.80±6.20	98.70±1.30	96.60±3.20	–	–	–	–	–	–	–	–	–	–	–	–
WTC	62.80±9.10	98.50±6.00	81.00±5.90	–	–	–	–	–	–	–	–	–	–	–	–
CLAM (NatBiomedEng'2021)	78.59±4.78	97.10±1.88	93.44±1.84	95.41±2.95	82.20±4.81	97.46±1.34	96.84±1.96	98.11±3.07	99.46±0.42	72.73±3.48	82.59±2.71	82.31±1.17	91.11±1.69	95.88±1.29	96.23±0.46
TransMIL (NeurIPS'2021)	69.17±6.59	94.11±1.99	95.11±1.74	96.56±2.16	96.54±1.57	98.77±0.87	98.66±1.29	96.54±2.69	99.58±0.27	74.60±3.68	87.55±3.39	84.38±1.13	90.28±1.44	95.04±1.15	95.63±0.57
DSMIL (CVPR'2021)	55.60±6.29	94.75±2.22	82.16±1.97	92.23±3.33	87.51±3.34	95.91±1.51	93.49±3.56	98.02±1.62	98.28±1.07	76.35±2.88	89.08±1.46	86.26±1.24	92.72±0.81	96.17±0.54	96.54±0.35
DTFD-MIL (CVPR'2022)	82.83±1.72	95.80±1.83	88.87±2.03	97.59±1.49	87.66±4.09	98.52±0.88	97.15±1.63	99.49±1.16	99.63±0.20	70.58±3.03	90.75±1.24	82.35±1.27	94.22±1.39	93.96±1.07	96.69±0.38
Bayes-MIL (ICLR'2023)	93.21±2.38	96.46±1.27	93.26±2.35	**98.79±1.52**	98.44±0.64	99.39±0.76	97.88±1.46	98.60±1.16	99.84±0.19	87.67±1.20	97.79±1.18	92.87±0.63	97.71±0.48	98.04±0.50	97.98±0.19
MambaMIL (MICCAI'2024)	92.98±2.69	98.35±1.06	94.92±1.83	97.88±1.78	99.04±0.84	99.32±0.68	98.48±1.53	98.60±1.40	99.66±0.49	87.97±1.17	**98.90±2.80**	93.62±0.84	96.26±0.55	97.42±0.43	97.96±0.18
RRTMIL (CVPR'2024)	90.46±2.50	98.86±0.32	95.63±1.74	97.88±1.78	98.33±1.39	99.32±0.59	98.58±2.75	98.70±1.42	99.77±0.31	88.97±1.36	97.56±0.81	92.76±0.65	97.85±0.44	97.28±0.76	98.47±0.21
WiKG (CVPR'2024)	93.21±2.38	98.76±0.91	96.67±0.91	95.17±2.76	83.65±3.95	95.03±2.11	99.51±1.10	95.57±2.18	99.48±0.32	86.31±3.56	98.21±0.62	92.16±2.14	97.16±0.73	97.32±0.86	98.69±0.40
Prov-GigaPath (Nature'2024)	96.05±1.54	98.06±0.66	97.21±0.80	97.73±1.40	98.46±2.45	99.54±0.26	99.19±1.25	99.03±1.25	99.89±0.12	79.07±3.91	94.46±0.78	93.15±1.53	**98.77±0.51**	97.06±0.46	98.80±0.30
2DMamba (CVPR'2025)	92.99±2.05	98.60±0.46	94.93±1.93	97.98±1.74	98.33±1.39	99.32±0.59	98.18±2.09	98.14±1.72	99.61±0.47	88.25±1.99	96.26±0.71	89.22±1.30	96.80±0.61	97.39±0.78	98.39±0.25
M3amba (CVPR'2025)	94.38±2.48	92.75±2.99	97.00±0.65	97.88±2.30	99.12±0.63	99.34±0.67	98.79±2.09	99.30±1.12	98.93±1.18	87.66±1.42	89.71±1.96	93.02±0.53	96.61±1.27	96.62±0.92	98.22±0.15
ABMILX (NeurIPS'2025)	92.80±2.32	98.89±0.26	95.94±1.37	97.27±2.16	98.44±0.74	99.08±0.38	98.48±1.53	98.60±1.40	99.75±0.33	85.36±4.19	89.52±5.26	93.70±0.81	97.46±0.44	97.43±0.38	98.07±0.19
C2MIL (ICCV'2025)	91.55±2.42	97.29±1.89	95.17±1.58	96.97±0.63	99.02±0.63	98.76±1.11	95.15±3.71	98.60±1.61	99.32±0.75	84.05±2.27	88.67±1.24	89.99±0.95	97.42±0.73	98.01±0.66	98.37±0.31
BioX-CPath (CVPR'2025)	97.24±2.14	96.78±1.16	98.17±2.78	98.17±2.78	99.44±0.54	98.33±1.34	99.49±1.16	99.67±0.23	76.76±4.74	88.46±0.65	86.47±4.32	97.65±0.37	98.38±0.42	98.58±2.75	98.62±0.23
WsiSSM(ours)	**96.67±0.91**	**99.05±1.05**	**97.44±0.90**	97.44±1.94	**99.24±1.22**	**99.64±0.28**	**99.47±0.80**	**99.53±0.72**	**99.95±0.05**	**89.30±2.64**	96.82±0.66	**93.79±1.31**	97.13±0.83	**98.39±0.91**	**98.87±0.40**

WsiSSM ranks first in Dice and IoU across five cohorts. All segmentation predictions are obtained without pixel-level training supervision (Table 2, Figure 6).

Table 2. Segmentation performance [%] of various MIL methods across multiple datasets.

Method	CAMELYON			PANDA			NCPDC			External A			External B		
	Accuracy	Dice	IOU	Accuracy	Dice	IOU	Accuracy	Dice	IOU	Accuracy	Dice	IOU	Accuracy	Dice	IOU
CLAM (NatBiomedEng'2021)	91.76±1.06	20.85±3.10	11.70±1.90	91.37±0.73	31.78±0.98	18.90±0.70	84.81±1.60	19.92±4.78	11.20±2.94	67.26±1.99	10.77±2.53	5.73±1.42	92.41±0.98	44.06±3.44	28.37±2.82
TransMIL (NeurIPS'2021)	97.55±0.79	18.69±5.22	10.47±3.24	91.18±0.64	31.01±1.55	18.37±1.08	92.95±3.08	6.47±1.90	3.36±1.03	88.97±2.64	27.01±7.17	15.95±4.58	92.08±1.29	22.32±1.70	12.90±1.09
DSMIL (CVPR'2021)	96.02±1.14	6.91±0.90	3.58±0.48	90.61±0.80	12.23±0.28	6.51±0.16	93.43±3.16	16.06±8.18	9.13±5.08	88.35±2.23	8.68±1.35	4.55±0.73	92.08±1.43	10.75±1.11	5.69±0.62
DTFD-MIL (CVPR'2022)	95.89±1.17	2.56±0.52	1.30±2.66	90.33±0.84	6.39±0.41	3.30±0.22	93.53±2.99	15.56±3.67	8.51±2.26	87.88±2.34	1.86±0.29	0.94±0.15	91.50±1.44	4.83±0.69	2.48±0.36
Bayes-MIL (ICLR'2023)	96.89±0.90	39.52±4.05	24.77±3.17	91.46±0.82	25.64±2.67	14.75±1.79	88.00±2.69	42.33±2.26	26.89±1.83	90.14±2.00	27.54±2.88	16.02±1.92	92.33±1.36	21.70±2.10	12.20±1.32
MambaMIL (MICCAI'2024)	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
RRTMIL (CVPR'2024)	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
WiKG (CVPR'2024)	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Prov-GigaPath (Nature'2024)	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
2DMamba (CVPR'2025)	96.60±1.05	27.29±3.94	15.90±2.62	90.89±0.80	16.60±1.39	9.06±0.83	86.09±3.27	24.85±3.61	14.27±2.39	89.20±2.16	21.21±2.22	11.89±1.40	92.80±1.25	28.58±2.03	16.70±1.37
M3amba (CVPR'2025)	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
ABMILX (NeurIPS'2025)	96.38±1.04	26.04±4.08	15.08±2.71	91.39±0.78	24.03±1.32	13.66±0.85	87.61±2.88	39.93±3.18	25.03±2.51	87.50±2.05	29.50±2.52	17.34±1.67	93.04±1.21	39.94±2.82	25.02±2.19
C2MIL (ICCV'2025)	96.66±1.09	30.46±6.52	18.27±4.53	90.98±1.49	24.16±1.74	13.76±1.12	87.41±3.12	41.87±3.47	26.58±2.76	86.67±2.38	10.60±3.90	5.68±2.18	91.64±1.27	27.22±5.28	15.95±3.61
BioX-CPath (CVPR'2025)	97.13±1.11	62.63±1.40	48.23±1.48	90.41±0.67	20.43±2.13	11.41±1.34	69.69±2.12	48.44±8.11	32.59±6.68	92.02±1.50	54.75±4.65	38.53±7.99	96.55±0.87	75.58±1.90	61.80±1.51
WsiSSM(ours)	**98.73±0.41**	**81.98±2.15**	**69.57±3.11**	**96.05±0.52**	**79.26±2.18**	**65.74±3.01**	**91.86±1.66**	**75.22±4.20**	**60.59±5.11**	**94.51±0.54**	**75.50±4.68**	**61.06±6.23**	**97.14±0.47**	**80.37±1.23**	**67.21±1.71**

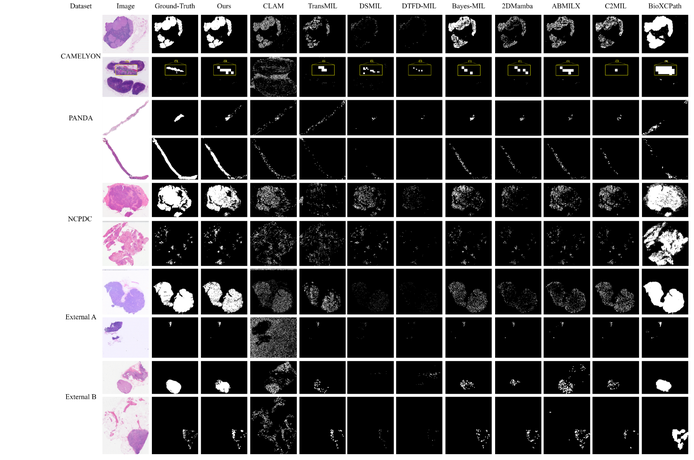


Figure 6. Visualization of segmentation mask from our and other MIL methods across multiple datasets.

Qualitative correspondence between predicted CAMs and matched IHC staining provides spatial corroboration of biomarker-associated localization, extending the framework beyond tumor–benign classification (Figure 7).

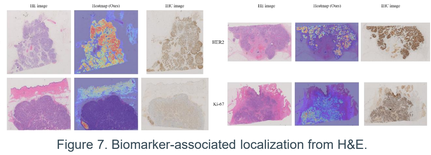


Figure 7. Biomarker-associated localization from H&E.

- Coordinate-preserving CAM reconstruction localizes sparse micro-lesions on a 128-μm patch grid, supporting small-focus identification within gigapixel tissue fields.(Figure 8).

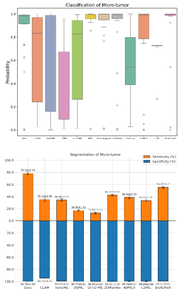


Figure 8. Micro-tumor evaluation.

- In simulated reading, AI assistance reduced pathologists' mean interpretation time by 8.03 s/slide and increased overall accuracy by 5.20 percentage points (+16.56 points for tumor slides), indicating subgroup-dependent assistance effects.

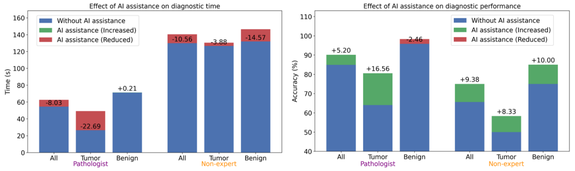


Figure 9. AI assistance in simulated diagnostic reading.

References :

[1] Shao, Z, et al, Transmil: Transformer based correlated multiple instance learning for whole slide image classification. NIPS 2021

[2] Child, R, et al, Generating long sequences with sparse transformers. arXiv 2019

[3] Samek, W , et al, Explaining deep neural networks and beyond: A review of methods and applications. Proceedings of the IEEE 109