SAM-Cluster and SAM-LP: Structure-Aware Evaluation Metrics for Selecting WSI Feature Extractors without MIL Training

Juhyeon Kim, Paul Ssemakula, Chanjae Song, and Mun Yong Yi†

Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea
{wedsed123, semaxspaul, chan4535, munyi}@kaist.ac.kr

Abstract. Recent foundation models for histopathology have created a rapidly expanding set of feature extractors (FEs), yet selecting an appropriate FE for a new whole-slide image (WSI) dataset typically requires repeatedly training Multiple Instance Learning (MIL) models, leading to substantial computational overhead. We propose an efficient pre-evaluation framework that ranks candidate FEs without MIL training. Existing representation evaluation metrics are largely designed for single-image inputs and fail to capture the bag-of-instances nature and spatial heterogeneity of WSIs. To address this mismatch, we introduce two WSI-aware metrics—SAM-Cluster and SAM-LP—that incorporate histological structure by leveraging the Segment Anything Model (SAM) to infer region-level organization and evaluate representation quality at the bag level. Experiments on eight public WSI datasets show that our metrics exhibit high rank correlation with the downstream performance of fully trained MIL models, enabling reliable FE selection at a fraction of the usual cost. Code is available at https://github.com/juhyeon-ai/WSI-FE-Selection.git

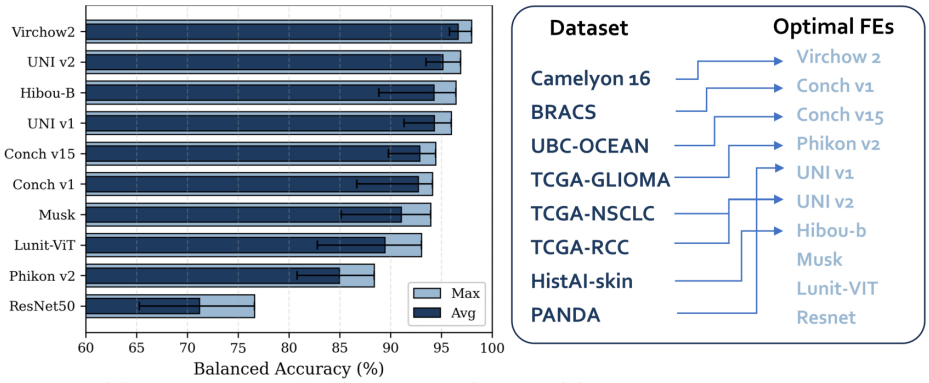
Keywords: Whole Slide Imaging · Multiple Instance Learning · Feature Extractor Selection · Segment Anything Model (SAM) · Foundation Models

1 Introduction

Whole Slide Image (WSI)-based computational pathology [3] has become a core technology for cancer diagnosis and prognosis. To process gigapixel-level images, the Multiple Instance Learning (MIL) framework [10, 21] is the standard approach. MIL partitions a single slide into thousands of patches (instances) to predict a slide-level (bag) label. Here, the Feature Extractor (FE) and the Aggregator form the two fundamental pillars determining overall model performance.

Recently, foundation models tailored for pathology imaging have rapidly proliferated. With public models like Virchow [29], UNI [4], and CONCH [16], researchers can leverage diverse FE options. However, previous studies have

† Corresponding author: munyi@kaist.ac.kr



(a) Impact of FE choice (Camelyon 16) (b) Dataset-specific Optimal FEs

Fig. 1. Performance comparison of FEs across MIL pipelines. (a) Performance of various FEs on the Camelyon16 dataset. (b) The highest-performing FE for each dataset.

predominantly focused on Aggregator architectures, leaving the impact of FE selection underexplored.

Through extensive preliminary experiments evaluating 10 foundation models and 9 MIL aggregators across 8 diverse datasets (detailed in Sec. 3.1), we found that performance varies significantly depending on the FE. For instance, on Camelyon16 (Fig. 1 (a)), an ImageNet-pretrained ResNet50 yielded 76.59% accuracy, whereas Virchow2 achieved 97.98%—a 20+ percentage point improvement. Furthermore, Fig. 1 (b) shows no universally optimal FE across all datasets: Virchow2 performed best on Camelyon16, Conch-v1 on BRACS, and UNI-v1 on PANDA. This result implies researchers must undergo costly trial-and-error to identify the optimal model per cohort. For instance, evaluating 10 FEs across 9 aggregators with 5 seeds requires 450 full MIL training runs per dataset—a prohibitive overhead that scales linearly with both the number of candidates and dataset size.

Dedicated methods for FE selection are absent in computational pathology. Existing pre-tuning evaluation metrics from general computer vision assume single-image inputs, failing to capture unique WSI properties. Supervised transferability metrics like LogME [27] and Linear Probing [13] require matching features with labels. In a MIL context, this requirement necessitates aggregating instance-level features (e.g., via Mean pooling) to match bag-level labels. This naive aggregation causes signal dilution, where sparse tumor features are overwhelmed by abundant normal patches. Conversely, unsupervised metrics like Self-Cluster [23] evaluate embedding quality without labels, avoiding dilution. However, by treating patches as independent and identically distributed (i.i.d.) samples, they disregard critical spatial heterogeneity and histological structures (e.g., tumor microenvironment). Thus, naively applying general vision metrics to WSIs is suboptimal.

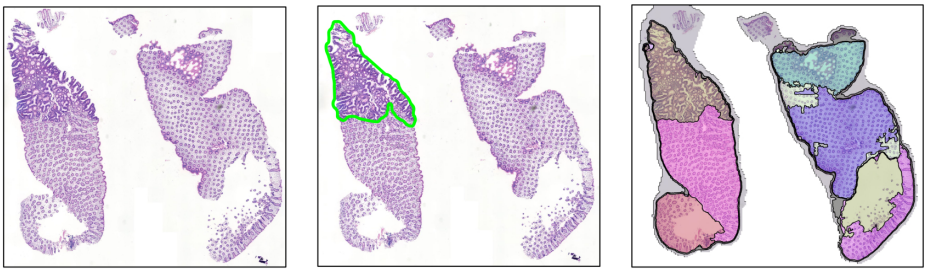


Fig. 2. Structural analysis of a colorectal WSI patch. (a) Original WSI. (b) Pathologist's annotation of a Tubulovillous Adenoma. (c) Zero-shot segmentation masks generated by SAM.

To overcome these limitations, we propose a structure-aware FE evaluation framework utilizing the Segment Anything Model (SAM) [12]. While SAM lacks specific medical semantic knowledge (e.g., identifying a Tubulovillous Adenoma), its zero-shot segmentation effectively captures inherent morphological boundaries that closely align with expert pathologist annotations (Fig. 2). Recent studies, such as SAM-MIL [6], also confirm that incorporating SAM's visual structural priors significantly enhances MIL representations. Leveraging this insight, we redesign existing metrics into SAM-Cluster and SAM-Linear Probing (SAM-LP) to evaluate intra-region consistency and inter-region separability based on meaningful regions of interest. Across eight cancer datasets, our method demonstrated high Spearman rank correlations with the downstream performance of fully trained MIL models. We confirmed that reliable FE selection is achievable using only a data subset. This capability elevates FE selection to a core design phase in computational pathology pipelines, significantly enhancing efficiency.

2 Methods

We formulate a 'Suitability Score' to quantitatively evaluate candidate Feature Extractor (FE) compatibility with a target WSI dataset prior to MIL training. Instead of exhaustive downstream training, our framework computes this score using pre-extracted features from a small randomly sampled WSI subset. Formally, given candidate FEs Φ, we seek the optimal ϕ^* maximizing the metric (Fig. 3 (a)):

$$\phi^* = \arg\max_{\phi \in \Phi} S_{\text{metric}}(\phi) \tag{1}$$

2.1 Preliminaries and Baseline Metrics

Let the i-th WSI be a bag of patches $X_i = \{x_{i,j}\}_{j=1}^{K_i}$ with label $Y_i \in \{1, \ldots, C\}$. The FE $\phi : \mathcal{X} \to \mathbb{R}^d$ maps each patch to an embedding $z_{i,j}$, and the Aggregator

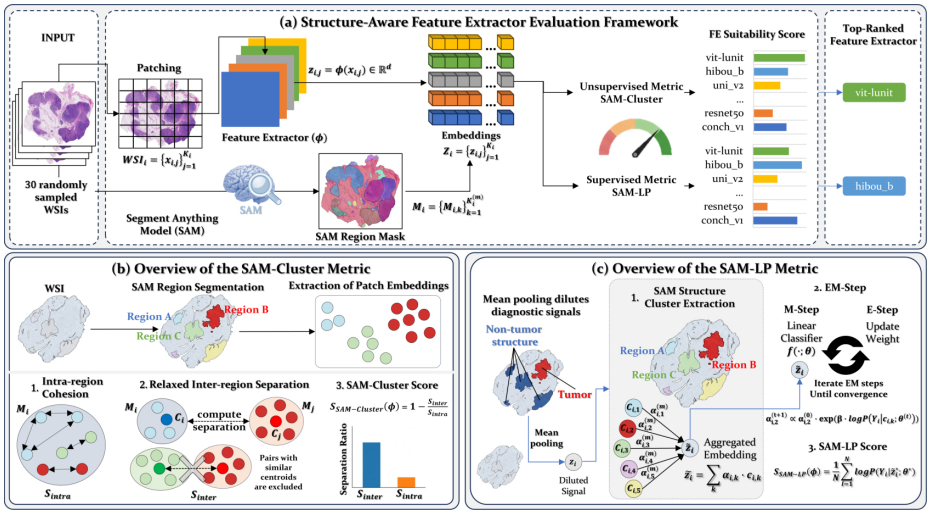


Fig. 3. Overview of the proposed structure-aware FE pre-evaluation framework. (a) The overall pipeline for ranking candidate FEs without full MIL training. (b) SAM-Cluster for evaluating unsupervised feature separability based on regional cohesion. (c) SAM-LP for assessing supervised task alignment via EM-based prototype optimization.

$\mathcal{A}$ predicts $\hat{Y}_i = \mathcal{A}(\{z_{i,j}\})$. An optimal FE should exhibit specific properties depending on label availability:

1. Unsupervised Settings. (1) *Feature Richness:* The feature space must not collapse. We use Effective Dimension (principal components explaining 99% variance) and NESum [8] (isotropy). (2) *Feature Separability:* Patches must be distinguishable. We use Self-Cluster (SC) [23] measuring global patch clustering. (Sec. 2.2 adapts this global metric to WSI structures via *SAM-Cluster*.)

2. Supervised Settings. To evaluate diagnostic task alignment, we utilize two baselines. *Linear Probing (LP)* [13] trains a linear classifier on the slide-level mean-pooled embedding $\bar{z}_i$. *LogME* [27] assesses feature-label compatibility by estimating maximum label evidence without gradient optimization. (Sec. 2.3 details how our approach mitigates the signal dilution limiting these methods.)

2.2 SAM-Cluster: Structure-Aware Unsupervised Separability

Global statistics-based separability (SC) inadequately captures WSI spatial heterogeneity. An ideal FE should maintain high consistency within the same tissue structure while establishing clear inter-tissue boundaries. To quantify this spatial consistency, we propose **SAM-Cluster**, utilizing Segment Anything Model (SAM) masks $\mathcal{M} = \{M_k\}$ as pseudo-structural labels (Fig. 3 (b)).

To mitigate SAM's potential over-segmentation, we introduce a relaxed separability measurement excluding mask pairs $\mathcal{E}$ with high centroid similarity. Let $c_k = \frac{1}{|Z^{(k)}|} \sum_{z \in Z^{(k)}} z$ be the mean vector of L_2-normalized patch embeddings

$Z^{(k)}$ in mask M_k. Given a threshold τ, the excluded set is $\mathcal{E} = \{(k,l) \mid c_k^\top c_l > \tau\}$. The intra-region cohesion (S_{intra}) and relaxed inter-region separability (S_{inter}) are:

$$S_{intra} = \mathbb{E}_k \left[\mathbb{E}_{z_i, z_j \in Z^{(k)}} \left[(z_i^\top z_j)^2 \right] \right] \tag{2}$$

$$S_{inter} = \mathbb{E}_{(k,l) \notin \mathcal{E}} \left[\mathbb{E}_{z_i \in Z^{(k)}, z_j \in Z^{(l)}} \left[(z_i^\top z_j)^2 \right] \right] \tag{3}$$

The SAM-Cluster Suitability Score is defined below, where higher values indicate superior preservation of histological boundaries:

$$S_{\text{SAM-Cluster}}(\phi) = 1 - \frac{S_{inter}}{S_{intra}} \tag{4}$$

2.3 SAM-LP: Structure-Aware Supervised Alignment

Conventional mean-pooling-based LP suffers from *Signal Dilution*: diagnostic signals from small regions (e.g., micrometastases) are overwhelmed by extensive background tissue. We propose **SAM-Linear Probing (SAM-LP)** to evaluate label alignment using structure-aware prototypes (Fig. 3 (c)).

Let $c_{i,k}$ be the L_2-normalized mask centroid of slide i. At iteration t, the slide embedding $\tilde{z}_i^{(t)}$ is reconstructed as a weighted sum of structural prototypes:

$$\tilde{z}_i^{(t)} = \sum_k \alpha_{i,k}^{(t)} c_{i,k}, \quad \text{s.t.} \quad \sum_k \alpha_{i,k}^{(t)} = 1 \tag{5}$$

Initial weights $\alpha_{i,k}^{(0)}$ are proportional to mask size. Since not all structures are diagnostically significant, we apply a single Expectation-Maximization (EM) update to refine the weights based on label relevance. In the **M-Step**, fixing the aggregated embedding $\tilde{z}_i^{(t)}$, we update the linear classifier parameters $\theta^{(t)}$ to maximize the average log-likelihood:

$$\theta^{(t)} = \arg\max_\theta \frac{1}{N} \sum_{i=1}^{N} \log P(Y_i \mid \tilde{z}_i^{(t)}; \theta) \tag{6}$$

In the **E-Step**, we update the weights based on each prototype's contribution to predicting Y_i, where β controls attention sharpness:

$$\alpha_{i,k}^{(t+1)} = \frac{\alpha_{i,k}^{(0)} \cdot \exp\left(\beta \log P(Y_i \mid c_{i,k}; \theta^{(t)})\right)}{\sum_j \alpha_{i,j}^{(0)} \cdot \exp\left(\beta \log P(Y_i \mid c_{i,j}; \theta^{(t)})\right)} \tag{7}$$

Following this update, the final Suitability Score evaluates downstream task alignment using the optimized embeddings $\tilde{z}_i^*$ and parameters θ^*:

$$S_{\text{SAM-LP}}(\phi) = \frac{1}{N} \sum_{i=1}^{N} \log P(Y_i \mid \tilde{z}_i^*; \theta^*) \tag{8}$$

3 Experiments

3.1 Experimental Setup

Ground Truth Definition. To establish a robust 'Ground Truth' (GT) for Feature Extractor (FE) suitability, we conducted exhaustive training. For each dataset, we fully trained 90 distinct configurations: 10 foundation models (Conch v1/v1.5 [16], Hibou-b [18], Lunit-vit [11], Musk [25], Phikon [7], ResNet [9], UNI v1/v2 [4], Virchow2 [29]) paired with 9 MIL aggregators (Mean pooling, AB-MIL [10], DS-MIL [14], CLAM [17], Trans-MIL [21], DTFD-MIL [28], WiKG [15], RRT-MIL [22], ILRA [26]). To ensure statistical reliability and minimize variance, every configuration was evaluated across five random seeds. The final GT ranking of an FE was determined by its average downstream performance across these runs.

Evaluation Protocol. The primary objective was to validate our pre-evaluation metrics. We assessed their reliability using Spearman's rank correlation (ρ) against the established GT rankings. For computational efficiency during metric evaluation, we randomly sampled subsets of 30 WSIs per dataset, repeating the process five times to maintain statistical stability via the Central Limit Theorem. Due to its significantly smaller image sizes, the PANDA dataset required sampling 20 times as many instances.

Datasets and Generalizability. To rigorously verify cross-domain generalizability, we benchmarked performance across eight public WSI datasets representing diverse organs and cancer types. Downstream performance was evaluated using Balanced Accuracy for seven datasets (Camelyon16 [5], BRACS [1], UBC-OCEAN [20], Histai-skin-b1 [19], TCGA-GLIOMA, TCGA-NSCLC, TCGA-RCC [24]). For PANDA [2], which involves ordinal prostate cancer grading, we utilized Quadratic Weighted Kappa (QWK).

3.2 Rank Correlation with Downstream MIL Performance

Table 1 presents the rank correlation between our methods and baselines, including four label-free unsupervised metrics and three supervised metrics.

Comparison of Unsupervised Metrics. SAM-Cluster averaged **0.5348**, outperforming Self-Cluster (0.4364) and NESum (0.4439). Notably, it consistently surpassed Self-Cluster across all 8 datasets ($p = 0.0078$, Wilcoxon signed-rank test). Higher correlations on structurally complex datasets (e.g., Camelyon16, BRACS, Histai-skin-b1) suggest that SAM-based consistency determines FE quality better than global statistics.

Comparison of Supervised Metrics. SAM-LP achieved the highest average predictive performance (**0.6258**), outperforming Linear Probing (0.5485). Although the overall improvement was not statistically significant ($p = 0.125$), a clear complexity-dependent trend was observed. On the structurally simplest datasets, PANDA (201 patches, 19.7 SAM regions) and TCGA-GLIOMA (2,994 patches, 53.2 SAM regions), LP slightly outperformed SAM-LP, while both methods achieved identical performance on UBC-OCEAN (3,175 patches, 70.5

Table 1. Spearman's rank correlation coefficients (ρ) between FE evaluation metrics and ground-truth MIL performance across 8 datasets. **Bold** indicates the best performance in each row.

Dataset	Unsupervised Metrics				Supervised Metrics		
	Eff. Dim	NESum	Self-Cluster	**SAM-Cluster**	LogME	Linear Prob.	**SAM-LP**
		[8]	[23]	(Ours)	[27]	[13]	(Ours)
Camelyon16	0.1030	0.3576	0.3697	**0.3939**	-0.1903	0.3091	0.3212
BRACS	-0.3939	0.0909	0.1273	0.3212	-0.2024	0.2364	**0.3697**
UBC-OCEAN	0.2242	0.6485	0.6121	0.6727	0.1806	**0.8303**	**0.8303**
TCGA-GLIOMA	0.3818	0.3455	0.1879	0.3697	**0.5103**	0.4545	0.4424
TCGA-NSCLC	-0.1636	0.4061	0.3212	0.4909	0.1976	0.6364	**0.9515**
TCGA-RCC	0.0303	0.5273	0.6485	0.7212	0.4836	0.6364	**0.7939**
Histai-skin-b1	0.3697	0.5879	0.6606	**0.7091**	0.1515	0.6727	**0.7091**
PANDA	0.5030	0.5879	0.5636	0.6000	**0.7430**	0.6121	0.5879
Average	0.1318	0.4439	0.4364	0.5348	0.2342	0.5485	**0.6258**

SAM regions). In contrast, SAM-LP consistently outperformed LP on all structurally more complex datasets, including BRACS (3,618 patches, 128.6 SAM regions), TCGA-NSCLC (3,834 patches, 90.3 SAM regions), and TCGA-RCC (4,000 patches, 71.9 SAM regions), where gains were substantial (e.g., TCGA-NSCLC: 0.6364 $\rightarrow$ **0.9515**). This progressive transition suggests that the benefit of structure-aware weighting increases with slide structural complexity.

Ground-Truth Robustness and Aggregator Bias. To verify aggregator-agnostic transferability, we analyzed the correlation between the averaged Ground Truth (GT) ranking and individual aggregator rankings. The averaged GT showed strong correlations with individual aggregator rankings (mean $\rho = 0.8205$) and the best-performing aggregator for each dataset ($\rho = 0.9299$), providing a representative estimate of FE suitability rather than reflecting specific aggregator bias. When stratified by aggregator, SAM-LP achieved the highest correlation with Mean pooling (0.6078) and RRT-MIL (0.5955), yielding an average correlation of 0.5688 ± 0.0226 across all nine aggregators. As expected, the linear-probing-based formulation of SAM-LP aligned most closely with Mean pooling. Meanwhile, SAM-Cluster achieved its highest correlation with CLAM (0.5260), although no clear architecture-specific pattern was observed.

3.3 Computational Efficiency and Robustness to Sample Size

Computational Efficiency. Exhaustive selection extracts features and trains M aggregators across S seeds for every $\phi \in \Phi$ over N slides, costing $C_{\text{exhaustive}} = N \times |\Phi| \times (K \times C_{\text{FE}} + M \times C_{\text{aggregator}})$ (where K is average patches/WSI). Our framework uses $n \ll N$ slides without aggregator training, costing $C_{\text{ours}} = n \times [|\Phi| \times (K \times C_{\text{FE}} + \alpha) + C_{\text{SAM}}]$ (α is metric overhead). On Camelyon16 (RTX A6000), typical extraction ($K \times C_{\text{FE}}$) takes $\approx$ 343 TFLOPs/WSI, while SAM (C_{SAM}) takes only 2.98 TFLOPs ($< 1\%$ of extraction), and α takes just 3.7 MFLOPs. Since SAM runs once per slide ($\mathcal{O}(1)$ w.r.t. $|\Phi|$), our cost scales by n rather than N, yielding a reduction ratio $\approx n/N$.

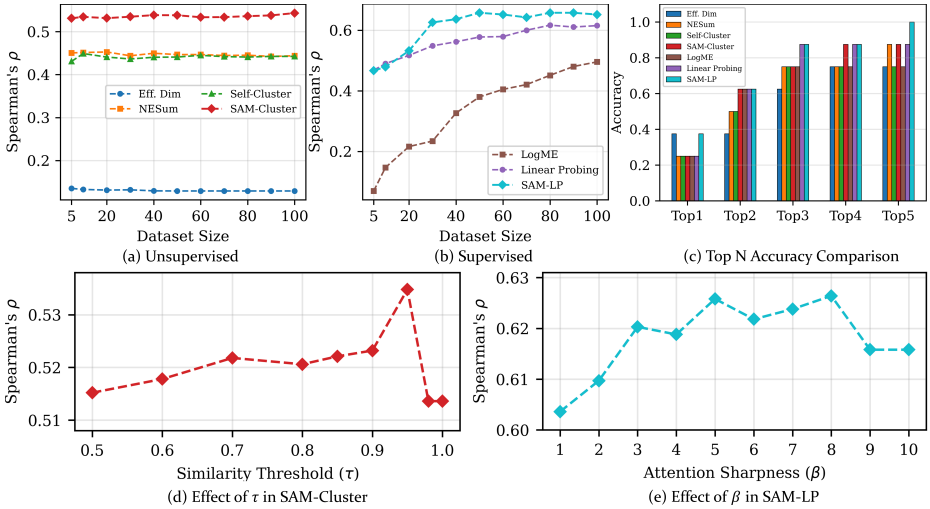


Fig. 4. Impact of sample size on rank correlation coefficients in (a) unsupervised and (b) supervised settings. (c) Top-k accuracy comparison for FE selection. Hyperparameter sensitivity analysis demonstrating the robustness of (d) SAM-Cluster to τ and (e) SAM-LP to β.

Robustness to Sample Size. For this efficiency to be practical, metrics must remain reliable on small subsets. We analyzed rank correlation across sample sizes $n \in [5, 100]$. In the unsupervised setting (Fig. 4 (a)), SAM-Cluster outperformed all baselines even at $n = 5$. In the supervised setting (Fig. 4 (b)), LP required $n \geq 80$ to peak, whereas SAM-LP approached its maximum at $n = 30$. SAM-LP's peak correlation ($\rho = 0.6576$) exceeded LP ($\rho = 0.6167$), guaranteeing higher reliability with substantially fewer slides.

3.4 Top-k Selection Analysis

The Top-k analysis (Fig. 4 (c)) demonstrates how this pre-evaluation reduces search costs in large-scale pipelines. Unsupervised SAM-Cluster exhibited stable screening capability (62.5%) even within a narrow Top-2 space. In the supervised setting, SAM-LP consistently included the optimal FE (100% hit rate) across all datasets when exploring the top 5 candidates.

3.5 Hyperparameter Sensitivity Analysis

We analyzed the sensitivity of our framework to key hyperparameters τ (Eq. (3)) and β (Eq. (7)) (Fig. 4 (d), (e)). Both metrics demonstrated high stability across broad parameter ranges, confirming that our methods do not require exhaustive dataset-specific tuning. We empirically set the similarity threshold $\tau = 0.95$ for SAM-Cluster and attention sharpness $\beta = 5$ for SAM-LP across all experiments.

4 Conclusion

This study pioneers preemptive FE evaluation for WSI datasets without prohibitively expensive full MIL training. We adapted representation metrics from general computer vision to the WSI context, systematizing baselines for both unsupervised and supervised settings. To overcome the spatial limitations of global statistical metrics, we proposed a structure-aware evaluation framework (SAM-Cluster, SAM-LP) leveraging SAM. Extensive experiments across 8 datasets demonstrated our metrics achieved the highest rank correlation with downstream MIL performance using minimal slide samples. Consequently, this research elevates FE selection from a costly trial-and-error phase to a theoretically justified core design step in large-scale computational pathology.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) under grant number RS-2022-NR068758. The authors also gratefully acknowledge the generous support provided by the Seegene Medical Foundation, Republic of Korea.

Disclosure of Interests

The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Brancati, N., Anniciello, A.M., Pati, P., Riccio, D., Scognamiglio, G., Jaume, G., De Pietro, G., Di Bonito, M., Foncubierta, A., Botti, G., et al.: BRACS: A dataset for breast carcinoma subtyping in H&E histology images. Database **2022**, baac093 (2022)
2. Bulten, W., Kartasalo, K., Chen, P.H.C., Ström, P., Pinckaers, H., Nagpal, K., Cai, Y., Steiner, D.F., Van Boven, H., Vink, R., et al.: Artificial intelligence for diagnosis and gleason grading of prostate cancer: The PANDA challenge. Nature Medicine **28**(1), 154–163 (2022)
3. Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., Fuchs, T.J.: Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature Medicine **25**(8), 1301–1309 (2019)
4. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., Shaban, M., et al.: Towards a general-purpose foundation model for computational pathology. Nature Medicine **30**(3), 850–862 (2024)
5. Ehteshami Bejnordi, B., Veta, M., Johannes van Diest, P., Van Ginneken, B., Karssemeijer, N., Litjens, G., Van Der Laak, J.A., CAMELYON16 consortium, Hermsen, M., Manson, Q.F., 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)

6. Fang, H., Huang, S., Tang, W., Huangfu, L., Liu, B.: SAM-MIL: A spatial contextual aware multiple instance learning approach for whole slide image classification. In: Proceedings of the ACM International Conference on Multimedia. pp. 6083–6092 (2024)
7. Filiot, A., Jacob, P., Mac Kain, A., Saillard, C.: Phikon-v2, a large and public feature extractor for biomarker prediction. arXiv preprint arXiv:2409.09173 (2024)
8. He, B., Ozay, M.: Exploring the gap between collapsed & whitened features in self-supervised learning. In: Proceedings of the International Conference on Machine Learning. pp. 8613–8634 (2022)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
10. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: Proceedings of the International Conference on Machine Learning. pp. 2127–2136 (2018)
11. Kang, M., Song, H., Park, S., Yoo, D., Pereira, S.: Benchmarking self-supervised learning on diverse pathology datasets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3344–3354 (2023)
12. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026 (2023)
13. Kornblith, S., Shlens, J., Le, Q.V.: Do better imagenet models transfer better? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2661–2671 (2019)
14. Li, B., Li, Y., Eliceiri, K.W.: Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14318–14328 (2021)
15. Li, J., Chen, Y., Chu, H., Sun, Q., Guan, T., Han, A., He, Y.: Dynamic graph representation with knowledge-aware attention for histopathology whole slide image analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11323–11332 (2024)
16. Lu, M.Y., Chen, B., Williamson, D.F., Chen, R.J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Le, L.P., Gerber, G., et al.: A visual-language foundation model for computational pathology. Nature Medicine **30**(3), 863–874 (2024)
17. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature Biomedical Engineering **5**(6), 555–570 (2021)
18. Nechaev, D., Pchelnikov, A., Ivanova, E.: Hibou: A family of foundational vision transformers for pathology. arXiv preprint arXiv:2406.05074 (2024)
19. Nechaev, D., Pchelnikov, A., Ivanova, E.: HISTAI: An open-source, large-scale whole slide image dataset for computational pathology (2025), https://arxiv.org/abs/2505.12120
20. Paayas, P., Annamalai, R.: OCEAN: Ovarian cancer subtype classification and outlier detection using DenseNet121. In: International Conference on Image Information Processing (ICIIP). pp. 827–831 (2023)
21. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al.: TransMIL: Transformer based correlated multiple instance learning for whole slide image classification. Advances in Neural Information Processing Systems **34**, 2136–2147 (2021)

22. Tang, W., Zhou, F., Huang, S., Zhu, X., Zhang, Y., Liu, B.: Feature re-embedding: Towards foundation model-level performance in computational pathology. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11343–11352 (2024)
23. Tsitsulin, A., Munkhoeva, M., Perozzi, B.: Unsupervised embedding quality evaluation. In: Topological, Algebraic and Geometric Learning Workshops 2023. pp. 169–188 (2023)
24. Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J.M.: The cancer genome atlas pan-cancer analysis project. Nature Genetics **45**(10), 1113–1120 (2013)
25. Xiang, J., Wang, X., Zhang, X., Xi, Y., Eweje, F., Chen, Y., Li, Y., Bergstrom, C., Gopaulchan, M., Kim, T., et al.: A vision–language foundation model for precision oncology. Nature **638**(8051), 769–778 (2025)
26. Xiang, J., Zhang, J.: Exploring low-rank property in multiple instance learning for whole slide image classification. In: Proceedings of the International Conference on Learning Representations (2023)
27. You, K., Liu, Y., Wang, J., Long, M.: LogME: Practical assessment of pre-trained models for transfer learning. In: Proceedings of the International Conference on Machine Learning. pp. 12133–12143 (2021)
28. Zhang, H., Meng, Y., Zhao, Y., Qiao, Y., Yang, X., Coupland, S.E., Zheng, Y.: DTFD-MIL: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18802–18812 (2022)
29. Zimmermann, E., Vorontsov, E., Viret, J., Casson, A., Zelechowski, M., Shaikovski, G., Tenenholtz, N., Hall, J., Klimstra, D., Yousfi, R., et al.: Virchow2: Scaling self-supervised mixed magnification models in pathology. arXiv preprint arXiv:2408.00738 (2024)