Positive Semi-definite Group-aware Instance Disentangled Learning for WSI Representation

Chentao Li[1], Behzad Bozorgtabar[2], Yifang Ping[3], Chun-Ka Wong[4], Pan Huang[5(✉)], and Jing Qin[5]

[1] Fu Foundation School of Engineering, Columbia University, New York, USA
[2] Aarhus University, Aarhus, Denmark
[3] Jingfeng Laboratory, Chongqing, China
[4] The University of Hong Kong, Hong Kong SAR, China
[5] Centre for Smart Health, School of Nursing, The Hong Kong Polytechnic University, Hong Kong SAR, China
✉ Corresponding author: Pan Huang (panhuang@polyu.edu.hk)

Abstract. Multiple instance learning (MIL) has been widely used for representing whole-slide pathology images. However, spatial, semantic, and decision entanglements among instances limit its representation and interpretability. To address these challenges, we propose a latent factor grouping-boosted cluster-reasoning instance disentangled learning framework for whole-slide image (WSI) interpretable representation in three phases. First, we introduce a novel positive semi-definite latent factor grouping that maps instances into a latent subspace, effectively mitigating spatial entanglement in MIL. To alleviate semantic entanglement, we employs instance probability counterfactual inference and optimization via cluster-reasoning instance disentangling. Finally, we employ a generalized linear weighted decision via instance effect re-weighting to address decision entanglement. Extensive experiments on multicentre datasets demonstrate that our model outperforms all state-of-the-art models. Moreover, it attains pathologist-aligned interpretability through disentangled representations and a transparent decision-making process. Code and datasets are available at github.com/Prince-Lee-PathAI/PG-CIDL

Keywords: Pathology · Whole-slide image · Multiple Instance Learning · Deep Clustering.

1 Introduction

Whole-slide images (WSIs), the gold standard for pathological classification, play a vital role in clinical tasks such as diagnosis, prognosis, and metastasis prediction [11]. Multiple instance learning (MIL) [18] has been promising approach for gigapixel WSIs analysis. Existing MIL frameworks suffer from weak interpretability because of entangled instance features from three aspects, *i.e.* spatial entanglement, semantic entanglement and decision entanglement; see Fig. 1.

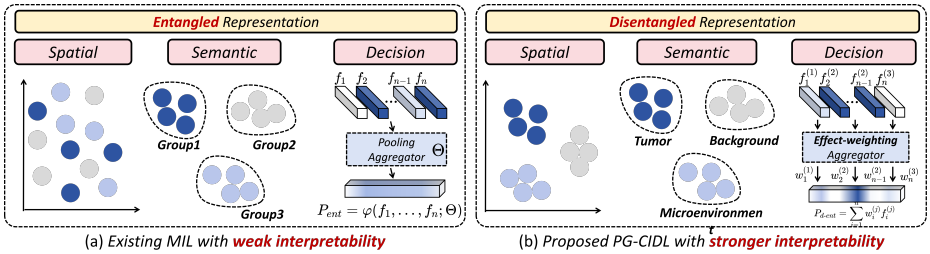

Fig. 1. Motivation of PG-CIDL. **(a):** Conventional MIL framework with entangled representation has weak interpretbility. **(b):** Proposed PG-CIDL framework with spatial, semantic and decision disentanglement has better interpretability.

Spatial entanglement arises when spatially adjacent instances exhibit correlated yet distinct pathological factors. It leads the model to confuse their individual contributions due to the absence of explicit spatial constraints. For example, the mixed tumor boundary and inflammatory areas make it difficult to distinguish the pathological border in spatial context. Typically, ABMIL [8] selects top-k patches as positive instances based on attention scores. ACMIL [25] extends this by introducing multi-branch attention and stochastic top-k masking to capture diverse informative instances. These attention-based MIL framework [12,9,3,25,8,18] generate attention scores that merely reflect the correlation between instances and bag-level predictions. They are incapable of modeling how instances contribute to pathological spatial patterns [17].

Despite spatial disentanglement, the inability to determine the factor group for each instance hinders the model's interpretability and representation. It is known that umor regions are more closely linked to pathological manifestations than non-tumor areas. However, conventional MIL models lack explicit semantic guidance, thus fusing tumor and non-tumor into unified latent representation. This semantic entanglement make the model fail to distinguish which regions are truly responsible for the diagnosis; see Fig. 1. Consequently, the learned instance weights fail to relate with pathological factors. To this end, cluster-incorporated MILs have been explored in WSI analysis. FuzzyMIL [15] decouples morphological patterns in WSIs through a learnable deep fuzzy clustering framework while cDPMIL [4] models the instance-to-bag structure using a cascade of Dirichlet processes based on feature covariance. These cluster-based models [16,10,19,15,4] merely approximate instance soft labels and suffer from entangled feature representation, which may highlight background noise and degrade both interpretability and classification performance.

With disentangled semantics, the ultimate decision-making can still be ambiguous. This decision entanglement results from the fuzzy inter-dependent aggregator in MIL, where instance contributions are determined by pooling or attention mechanisms; see Fig. 1. Hence, the importance of each patch is affected by the global feature distribution, blurring the relationship between attention

weights and true pathological relevance. DGR-MIL [26] utilizes a set of global vectors to represent WSI features, while RRT-MIL [21] re-embeds degraded features via a regional transformer. Causality-based MIL models have also been explored to capture instance correlations [22,6,14,13,2]. For example, MFC-MIL [6] employs an adaptive memory module to simulates causal interventions to mitigate confounders in diagnosis tasks. In general, although these methods tend to refine WSI representation, not only the final decision-making remains ambiguous without specific causal effect weighting, but they lack structural causal models (SCMs) to build disentangled relationships.

To address these entanglement, we introduce PG-CIDL, a three-phase disentangled representation learning framework for interpretable WSI analysis, which enables the model to disentangle spatially, semantically and decision ambiguous representations; see Fig. 1. First, we proposed a novel positive semi-definite latent factor grouping (PSD-LFG) method, which maps spatially entangled instances into a implicit subspace and adaptively partition them into three latent groups. To address semantic entanglement, we develop the cluster-reasoning instance disentangling (CID) in the second phase. By instance probability counterfactual inference, we measure its causal effect, for which features are disentangled by the extent of effects. Finally, the obtained effects are used to re-weighting original feature to get a refined WSI representation via generalized linear weighted summation, mitigating decision entanglement.

2 Methodology

2.1 Problem Statement

Assumption: (1): Tumor, microenvironment, and irrelevant background can coexist within each WSI. These weakly-supervised patches meet the criteria for adopting the MIL framework. **(2)**: Factors hold causal dependencies in disentangled representation learning (DRL), rather than independence. **(3)**: Tumor cells are directly related with pathological grading, while microenvironment interacts with tumor cells, jointly shaping the grading outcome.

MIL Formulation: The goal of WSI-based classification task in MIL framework is to learn a permutation-invariant scoring function $\mathcal{F}$ that maps the set of instances $\mathbf{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ to label $\mathbf{Y}$. The prediction process can be formulated as follows:

$$\hat{\mathbf{Y}} = \mathcal{F}(\mathbf{X}) \quad \text{where} \quad \mathcal{F}(\mathbf{X}) \equiv f\big(\sigma(\mathcal{A}(\mathbf{X}))\big) \tag{1}$$

where $\sigma(\cdot), f(\cdot)$ denote the aggregation function, prediction head with softmax normalization, respectively. $\mathcal{A}(\cdot)$ denotes the feature extractor.

DRL Formulation: We formulate DRL with SCMs, optimization problems and proof for each phase respectively in Supplementary Material.

2.2 Positive Semi-definite Latent Factor Grouping

Classic clustering methods with L_p norm with equal weights disregard the interaction and fusion between each two dimensions, exacerbating the spatial entan-

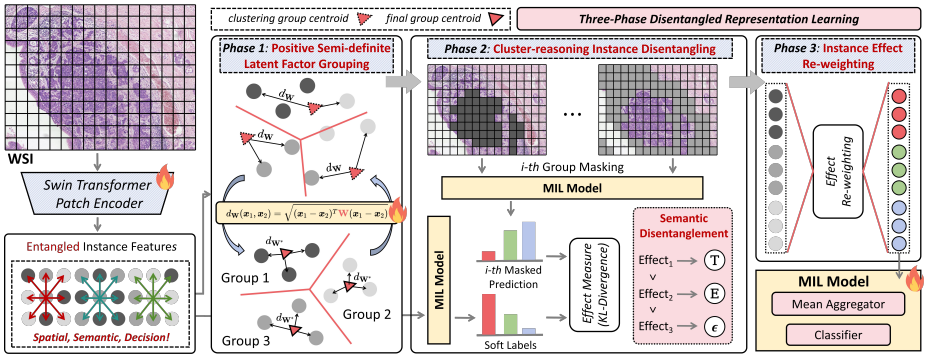

Fig. 2. Overview of **PG-CIDL**. Entangled Features extracted by encoder are categorized into three groups via **P**ositive semi-definite latent factor **G**rouping. They are identified into tumor, microenvironment and background factors by **C**luster-reasoning **I**nstance **D**isentangling. End-to-end optimization is performed across all phases for disentangled representation **L**earning with instance effect re-weighted features.

glement and semantic ambiguity. Inspired by the idea of metric learning and deep clustering [20,1], we proposed a latent factor grouping approach with positive semi-definite constraint to address the limitations.

The general form of metric learning is based on Mahalanobis distance.

$$d_{\mathbf{W}}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sqrt{(\boldsymbol{x}_1 - \boldsymbol{x}_2)^\top \mathbf{W} (\boldsymbol{x}_1 - \boldsymbol{x}_2)} \tag{2}$$

where $\mathbf{W} \in \mathbb{R}^{n \times n}$ is a trainable positive semi-definite matrix and $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^{n \times 1}$ denote the instance feature representation vectors.

Positive semi-definite matrix ensures the legality and non-negativity of the metric and facilitates robust metric learning process. Furthermore, compared to diagonal matrix, a general positive semi-definite matrix models more interaction effects among pathological factors, while a diagonal matrix only adjusts the weight of each feature independently. For any $\mathbf{A} \in \mathbb{R}^{n \times n}$, the matrix $\mathbf{A}^\top \mathbf{A}$ is always positive semi-definite. Therefore, we replace the L_p norm with the adaptive distance $d_{\mathbf{W}}(\cdot, \cdot)$ and ensure $\mathbf{W} \succeq 0$ by parameterizing $\mathbf{W} = \mathbf{A}^\top \mathbf{A}$, $\mathbf{A} \in \mathbb{R}^{r \times n}$, where $r < n$ ensures a low-rank structure for easier computation. By incorporating optimal $\mathbf{W}^* = \mathbf{A}^{*\top} \mathbf{A}^*$ into KMeans ($k = 3$, Fig. 2) we can also assume that PSD-LFG maps the distance into a latent subspace in (3), where we can perform more robust and accurate clustering results, thus alleviating spatial entanglement.

$$d_{\mathbf{W}^*}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sqrt{(\boldsymbol{x}_1 - \boldsymbol{x}_2)^\top \mathbf{A}^{*\top} \mathbf{A}^* (\boldsymbol{x}_1 - \boldsymbol{x}_2)} = \|\mathbf{A}^* (\boldsymbol{x}_1 - \boldsymbol{x}_2)\|_2 \tag{3}$$

2.3 Cluster-reasoning Instance Disentangling

Based on the results from Phase 1, we obtained three latent yet semantically entangled instance groups. Therefore, we proposed CID, which leverages counterfactual inference to estimate the causal effects within factors, thereby disentangling ambiguous semantics. Fig. 2 and Algorithm in Supplementary Material intuitively present the idea of CID, which involves the following three steps.
Group Masking: Let $\mathcal{A}(\mathbf{X}) = \mathbf{Z} \rightarrow \{\mathbf{Z}_1, \mathbf{Z}_2, \mathbf{Z}_3\}$ be three instance groups from phase 1. The intervention sets $\mathsf{do}(\mathbf{Z}_{\setminus k})$ as $\mathbf{Z}_{\setminus k} = \mathbf{Z} \setminus \mathbf{Z}_k, k = 1, 2, 3$. This masking severs all back-door paths through $\mathbf{Z}_k$ and thus isolates its causal effect. Then the prediction under that intervention is $P_k = f(\sigma(\mathbf{Z}|\mathsf{do}(\mathbf{Z}_{\setminus k})))$, where we consider the counterfactual of factor $\mathbf{Z}_k$ as $\mathbf{Z}_{\setminus k}$ that masking $\mathbf{Z}_k$. This masked prediction serve as an important component for causal effect measurement.
Effect Measurement: Let $P_v = f(\sigma(\mathbf{Z})$ be the vanilla prediction without any intervention, which serves as the reference and soft labels for comparison. We regard P_v as the target distribution and P_k as the ones approximate the target. Due to this asymmetry, we employ Kullback-Leibler divergence to calculate pairwise difference D_k as effects in (4).

$$D_k = D_{KL}(P_v||P_k) = \sum_z P_v(z) \log(\frac{P_v(z)}{P_k(z)}) \tag{4}$$

where D_k reveals how much the masked group influence the original prediction.
Semantic Disentangling: A larger D_k suggests a more pathologically important group (e.g. tumor), while a smaller one indicates irrelevant groups (e.g. background and microenvironment). Consequently, we can causally identify the *factors of variation* of each group by tumor, microenvironment and background; see Fig. 2 and Algorithm in the Supplementary Material.

2.4 Instance Effect Re-weighting and Optimization

We normalize each "leave-one-out" prediction divergence D_k to get approximate effect $w_k = \frac{D_k}{\sum_j D_j}$ and perform effects re-weighting to refine the representation. Here $\mathbf{Z}_{\text{final}} = \texttt{mean}\big(\sum_{k=1}^{3} w_k \, \mathbf{Z}_k\big)$can reflect how the previously disentangled instances effect the final WSI representation and mitigating decision entanglement. Finally, we perform end-to-end optimization across all stages to learn more task-related, interpretable and generalized feature representations. Specifically, the loss function is formulated as follows:

$$\begin{aligned} \mathcal{L} =& \mathcal{L}_{ce}(f(\sigma(\mathbf{Z}), \mathbf{y})) + \mathcal{L}_{ce}(f(\sigma(\mathbf{Z}_{\text{final}}), \mathbf{y}')) + \lambda \cdot d(\mathbf{Z}, \mathbf{Z}_{\text{final}}) \\ & -\gamma_1 \cdot d_{\text{reg}} + \gamma_2 \cdot d_{\text{compact}} \end{aligned} \tag{5}$$

where $\mathbf{y}'$ takes the counterfactual pseudo label from original output probability distribution $f(\sigma(\mathbf{Z})$ by using the second maximum logits. The second and third terms are to generate minimally perturbed counterfactual representations while successfully inducing a counterfactual label transition.

Table 1. Performance comparison of WSI classification on multicentre datasets with Swin-T pretrained features.

Models	AMU-CSCC		AMU-LSCC		CAMELYON16		DHMC-LUNG	
	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
ABMIL [8]	0.8302	0.9547	0.6739	0.7888	0.7429	0.7918	0.7333	0.8123
CLAM-SB[16]	0.8302	0.9581	0.6957	0.8367	0.7429	0.7478	0.7556	0.7783
CLAM-MB[16]	0.8774	0.9336	0.6812	0.8031	0.7429	0.7339	0.7333	0.7684
TransMIL [18]	0.8868	0.9521	0.8261	0.9021	0.7429	0.6898	0.7333	0.8084
DTFD-MIL [24]	0.8113	0.9036	0.5435	0.7245	0.7143	0.7306	0.6889	0.6983
IBMIL-DS [13]	0.7830	0.8881	0.6377	0.7672	0.7429	0.7037	0.6889	0.7091
ILRA-MIL[23]	0.8491	0.9123	0.6884	0.8048	0.7714	0.7306	0.7778	0.8074
S4MIL [7]	0.8491	0.9605	0.7971	0.8845	0.7429	0.7322	0.7111	0.7007
FRMIL [5]	0.8113	0.8990	0.6377	0.8020	0.7143	0.6914	0.7111	0.7442
DGR-MIL [26]	0.8868	0.9613	0.8188	0.9147	0.8286	0.8253	0.8000	0.7886
ACMIL-MHA [25]	0.9057	0.9701	0.8551	0.9219	0.8000	0.7331	0.6889	0.6849
RRT-MIL [21]	0.8585	0.9265	0.7681	0.8732	0.7143	0.7608	0.7778	0.7521
MFC-MIL [6]	0.8962	0.9723	0.8261	0.9126	0.7714	0.7208	0.7778	0.7931
PG-CIDL (ours)	**0.9434**	**0.9859**	**0.9275**	**0.9719**	**0.8857**	**0.9282**	**0.8444**	**0.8820**

The last two terms are to enlarges the distance among causally influential clusters and minimizes intra-cluster variance, respectively. They can thus together improve the assignment of different instances by optimizing the trainable matrix $\mathbf{W}$ by computing the gradients $\frac{\partial \mathcal{L}}{\partial \mathbf{A}}$.

$$
\begin{aligned}
d_{\text{reg}} &= d_{\mathbf{W}}\Big(\text{mean}(\mathbf{Z}^{\text{TC}}), \text{mean}(\mathbf{Z} \setminus \mathbf{Z}^{\text{TC}})\Big) \\
d_{\text{compact}} &= \frac{1}{K}\sum_{k=1}^{K}\frac{1}{|C_k|}\sum_{x_i \in C_k}\|x_i - \mu_k\|_2
\end{aligned}
\tag{6}
$$

where μ_k is the centroid of each cluster.

3 Experimental Results

3.1 Datasets and Setup

To validate PG-CIDL for diagnosis and sub-typing tasks, we conduct extensive experiments on two public datasets (CAEMYON16, DHMC-LUNG) and two private datasets from Army Medical University, where **AMU-LSCC** is for laryngeal squamous cell carcinoma pathological grading and **AMU-CSCC** is for cervical squamous cell carcinoma. AMU-LSCC dataset includes 342 whole slide images that was divided by a 6:4 training-validation ratio for each category. Specially Grade I contains a total of 89 WSIs, Grade II contains 152 WSIs and Grade III includes 101 WSIs. AMU-CSCC dataset includes 262 whole slide images, where Grade I contains a total of 27 WSIs, Grade II contains 127 WSIs and Grade III includes 108 WSIs.

Table 2. Ablation experiments on different clustering methods in PG-CIDL. The baseline model is vanilla Swin Transformer.

Models	AMU-CSCC		CAMELYON16	
	ACC	AUC	ACC	AUC
MFC-MIL	0.8962	<u>0.9723</u>	0.7714	0.7208
ACMIL-MHA	<u>0.9057</u>	0.9701	<u>0.8000</u>	<u>0.7331</u>
baseline	<u>0.8868</u>	0.9778 ↑	0.8000↑	0.8588↑
PG-CIDL w/o PSD-LFG	0.8962 ↑	0.9641 ↓	0.8286	0.8971
PG-CIDL w/o CID	0.8868	0.9825	0.8571	0.8865
PG-CIDL	**0.9434** ↑	**0.9859** ↑	**0.8857**↑	**0.9282**↑

Table 3. Ablation experiments on different metrics in CID.

Metrics in CID	AMU-CSCC		CAMELYON16	
	ACC	AUC	ACC	AUC
Cosine distance	0.9151	0.9776	0.6857	0.7771
Hellinger distance	0.9245	0.9795	<u>0.8857</u>	<u>0.9151</u>
JS divergence	<u>0.9434</u>	<u>0.9849</u>	0.8571	0.8882
KL divergence (ours)	**0.9434**	**0.9859**	**0.8857**	**0.9282**

3.2 Comparison with SOTA Methods

Table 1 presents the performance of SOTA methods on multicentre datasets. The results suggest that our PG-CIDL outperforms all SOTA models across all three evaluation metrics and four benchmarks. For features pretrained on ImageNet-11k, while ACMIL-MHA achieves the second-best classification performance on AMU-LSCC, our model improves the ACC and AUC by 7.24% and 5.00%, respectively. On CAMELYON16 dataset, the improvement is 5.71% and 10.29% compared to DGR-MIL. AMU-CSCC and DHMC-LUNG have also seen the similar trend that our proposed PG-CIDL achieves the best classification performance. We attribute this achievement to the advantages of refined representation from PSD-LFG and CID in our disentangled representation learning framework, which will be validated in the following sections.

3.3 Ablation Experiments

To validate the effectiveness of components proposed in PG-CIDL, we conducted extensive ablation experiments. In Table 2, we validate PSD-LFG by replacing with L2-KMeans, we observe a 0.94% improvement in ACC and a 1.37% drop in AUC. Incorporating PSD-LFG, PG-CIDL leads to optimal classification performance. With fixed attention, our PG-CIDL degrades without causal effects re-weighting from CID. Instead, combining these two components improves both ACC and AUC. Public datasets like CAMELYON16 also implies effectiveness of CID and PSD-LFG modules in PG-CIDL by ACC and AUC compared to baseline.

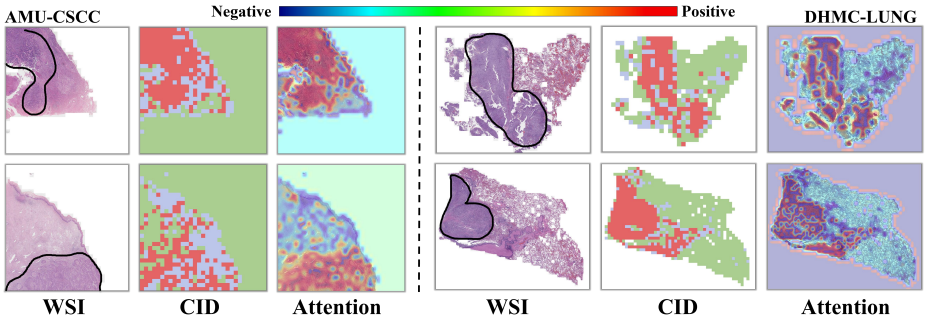

Fig. 3. Visualization on AMU-CSCC and DHMC-LUNG. The first column is tumor ground truth. The second is the grouping result, where red, blue, green denote tumor, environment and background respectively. The last is the attention heatmap.

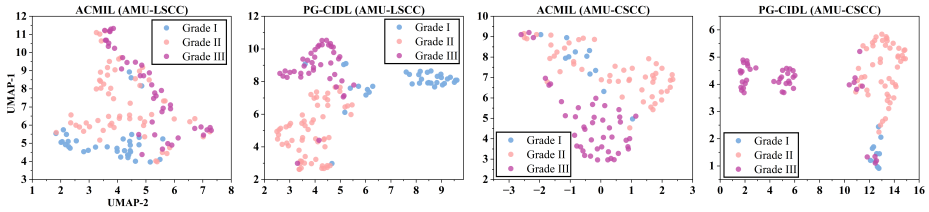

Fig. 4. 2-D feature representation distribution on multicentre datasets.

Moreover, Table 3 reports the performance of different metrics employed in CID module. Cosine distance is the least suitable metric. As a symmetric version of KL, the JS divergence achieves comparable performance with only a 0.1% drop in ACC on the AMU-CSCC. Overall, these results demonstrate that incorporating KL divergence in CID module yields superior performance.

3.4 Visualization of WSI Interpretability and Representation

To validate whether our model managed to identity instance features into three disentangled factors, we provide visualization results of our PG-CIDL on AMU-CSCC, AMU-LSCC and DMHC-LUNG. In Fig. 3 and Supplementary Material, we can see that both disentangled factors and attention heatmaps highlight the area that are highly consistent with pathologists' annotations, mirroring a real-world diagnostic workflow. Therefore, it demonstrates that PG-CIDL not only learns more interpretable and task-relevant representations, but also aligns well with clinical reasoning, offering a comprehensive explanation of predictions.

Moreover, we utilize UMAP to visualize feature representation in 2-D space. In Fig. 4, our proposed PG-CIDL show less entanglement within feature distributions, while other sub-optimal models without spatial disentanglement show

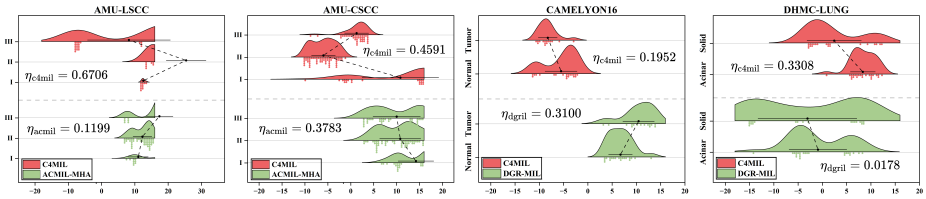

Fig. 5. ANOVA plots on four datasets where η_* denote the value of effect size.

more overlaps. We also conduct the analysis of variance (ANOVA) in Fig. 5 with effect size metrics (η^2). A larger effect size suggests a stronger ability of the model to distinguish differences between classes. In Fig. 5, PG-CIDL models the distinct distributions within different categories, while ACMIL and DGR-MIL tend to overlap. It has larger effect size 0.6706, 0.4591, 0.3308 compared to the second-best models on AMU-CSCC, AMU-LSCC and DHMC-LUNG, respectively. It not only implies that PG-CIDL disentangles feature spatially, but it also has superior ability to distinguish features related to corresponding grades.

4 Conclusion

WSI-based pathological grading is the gold standard for clinical diagnosis and prognosis. However, mainstream multi-instance learning frameworks for WSI classification often suffer from limited interpretability and generalizability due to spatial, semantic, and decision entanglement. Thus, we propose PG-CIDL, a three-phase disentangled representation learning framework that integrates latent factor grouping, cluster-reasoning instance disentangling, and instance effect re-weighting. PG-CIDL decomposes entangled features into three causally relevant factors, i.e. tumor, microenvironment, and background, yielding more interpretable WSI representations. Extensive experiments on multicentre datasets demonstrate superior classification performance, strong interpretability, and generalizability.

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

References

1. Caron, M., Bojanowski, P., Joulin, A., Douze, M.: Deep clustering for unsupervised learning of visual features. In: Proceedings of the European conference on computer vision (ECCV). pp. 132–149 (2018)
2. Chen, K., Sun, S., Zhao, J.: Camil: Causal multiple instance learning for whole slide image classification. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 1120–1128 (2024)

3. Chen, R.J., Lu, M.Y., Weng, W.H., Chen, T.Y., Williamson, D.F., Manz, T., Shady, M., Mahmood, F.: Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4025 (2021)
4. Chen, Y., Chan, T.H., Yin, G., Jiang, Y., Yu, L.: cdp-mil: Robust multiple instance learning via cascaded dirichlet process. In: European conference on computer vision. pp. 232–250. Springer (2024)
5. Chikontwe, P., Kim, M., Jeong, J., Sung, H.J., Go, H., Nam, S.J., Park, S.H.: Frmil: Distribution re-calibration based multiple instance learning with transformer for whole slide image classification. IEEE Trans. Med. Imaging (2024)
6. Cui, X., Chen, W., Su, J.: A multiscale frequency domain causal framework for enhanced pathological analysis. In: The Thirteenth International Conference on Learning Representations (2025)
7. Fillioux, L., Boyd, J., Vakalopoulou, M., Cournède, P.H., Christodoulidis, S.: Structured state space models for multiple instance learning in digital pathology. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 594–604. Springer (2023)
8. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International Conference on Machine Learning. pp. 2127–2136. PMLR (2018)
9. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning. pp. 2127–2136. PMLR (2018)
10. Jin, J., Wang, S., Dong, Z., Liu, X., Zhu, E.: Deep incomplete multi-view clustering with cross-view partial sample and prototype alignment. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11600–11609 (2023)
11. Kumar, N., Gupta, R., Gupta, S.: Whole slide imaging (wsi) in pathology: current perspectives and future directions. Journal of digital imaging **33**(4), 1034–1040 (2020)
12. 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)
13. Lin, T., Yu, Z., Hu, H., Xu, Y., Chen, C.W.: Interventional bag multi-instance learning on whole-slide pathological images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19830–19839 (2023)
14. Lin, W., Zhuang, Z., Yu, L., Wang, L.: Boosting multiple instance learning models for whole slide image classification: A model-agnostic framework based on counterfactual inference. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 3477–3485 (2024)
15. Liu, A., Li, T., Cai, J., Vajrala, S.S.V.: Fuzzymil: Decoupling pathological phenotypes through deep fuzzy clustering for efficient whole slide image analysis. In: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2025)
16. 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. Nat. Biomed. Eng. **5**(6), 555–570 (2021)
17. Pawlowski, N., Coelho de Castro, D., Glocker, B.: Deep structural causal models for tractable counterfactual inference. Advances in neural information processing systems **33**, 857–869 (2020)

18. 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)
19. Song, A.H., Chen, R.J., Ding, T., Williamson, D.F., Jaume, G., Mahmood, F.: Morphological prototyping for unsupervised slide representation learning in computational pathology. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11566–11578 (2024)
20. Tang, H., Chen, K., Jia, K.: Unsupervised domain adaptation via structurally regularized deep clustering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8725–8735 (2020)
21. 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)
22. Wu, X., Wang, H., Wu, H.: Causal attention multiple instance learning for whole slide image classification. In: The 16th Asian Conference on Machine Learning (Conference Track) (2024)
23. Xiang, J., Zhang, J.: Exploring low-rank property in multiple instance learning for whole slide image classification. In: The Eleventh International Conference on Learning Representations (2023)
24. 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)
25. Zhang, Y., Li, H., Sun, Y., Zheng, S., Zhu, C., Yang, L.: Attention-challenging multiple instance learning for whole slide image classification. In: European Conference on Computer Vision. pp. 125–143. Springer (2024)
26. Zhu, W., Chen, X., Qiu, P., Sotiras, A., Razi, A., Wang, Y.: Dgr-mil: Exploring diverse global representation in multiple instance learning for whole slide image classification. In: European Conference on Computer Vision. pp. 333–351. Springer (2024)