Supplementary Material

Anonymized Authors

Anonymized Affiliations
`email@anonymized.com`

1 Problem Formulation

WSI-based MIL framework suffer from weak interpretability because of spatial, semantic and decision entanglement. Therefore, we apply DRL to address these entangled dependencies [5].

DRL Formulation: Given the set of instances $\mathbf{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, we assume factor tumor ($\mathbf{T}$), microenvironment ($\mathbf{E}$) and background noise (ϵ) ambiguously generates the whole distribution by function $\boldsymbol{\Phi}(\cdot)$. Thus, DRL aims to disentangle this mapping from given instances to hidden factors $\mathbf{F} = \{\mathbf{h}_1 := \mathbf{T}, \mathbf{h}_2 := \mathbf{E}, \mathbf{h}_3 := \epsilon\}$, *i.e.*,

$$\mathbf{X} = \boldsymbol{\Phi}(\mathbf{F}) \quad \rightarrow \quad \mathbf{h}_j = \boldsymbol{\Phi}^{-1}(\boldsymbol{x}_i) \tag{1}$$

1.1 Spatial Disentanglement

In this case, spatial disentanglement can be modeled by expanding the distance $\mathcal{D}$ between each two factors, *i.e.*, maximizing the following,

$$\max_{\mathcal{A},\boldsymbol{\Phi}} \mathcal{D} = \sum_{j \neq j'} \left\| \mathbb{E}_{\boldsymbol{x} \sim \boldsymbol{\mathcal{X}}} \left[\mathbb{1}_{\mathbf{h}_j} \left(\boldsymbol{\Phi}^{-1}(\boldsymbol{x}) \right) \mathcal{A}(\boldsymbol{x}) \right] - \mathbb{E}_{\boldsymbol{x}' \sim \boldsymbol{\mathcal{X}}} \left[\mathbb{1}_{\mathbf{h}_{j'}} \left(\boldsymbol{\Phi}^{-1}(\boldsymbol{x}') \right) \mathcal{A}(\boldsymbol{x}') \right] \right\|_2 \tag{2}$$

1.2 Semantic Disentanglement

By the assumption, *factors of variation* in pathological grading are not independent semantics but hold certain causal relations. To address semantic entanglement, we introduce two decoupled SCMs to characterizes these disentangling factors in DRL as prior knowledge. In Figure 1, both two SCMs meet the backdoor criterion, which provides a theory for using Pearl's back-door adjustment to estimate the causal effect of one factor by controlling another; see (3). Thus, it guarantees the feasibility of PG-CIDL model desgin.

$$P\big(\mathbf{G} \mid \mathsf{do}(\mathbf{T})\big) = \sum_k P\big(\mathbf{G} \mid \mathbf{T}, \mathbf{E} = e_k\big) P(\mathbf{E} = e_k) \tag{3}$$

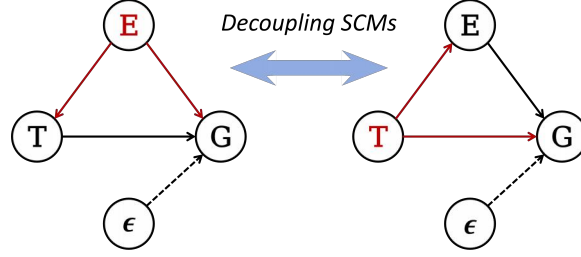where e_k denotes the controlling option of k instances belonging to factor $\mathbf{E}$.

Fig. 1. Two decoupled SCMs for disentangling factors in DRL, where node T is for factor tumor, node E for microenvironment, node G for pathological grading outcome and node ϵ for possible background noise.

1.3   Decision Disentanglement

Furthermore, we employ information entropy to determine the decision disentanglement. Let $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be random variables of instance in one bag. Then the following theorem holds [4]:

Theorem 1. *The joint information entropy can be expressed as $\sum_{i=1}^{n} H(\mathbf{x}_i)$ iff. $\mathbf{x}_i$ are i.d.d. variables. Furthermore, when they are not i.d.d., we can derive:*

$$H(\mathbf{x}_1, \cdots, \mathbf{x}_n) = H(\mathbf{x}_1) + \sum_{i=2}^{n} H(\mathbf{x}_i|\boldsymbol{\Phi}^{-1}(\mathbf{x}_1, \cdots, \mathbf{x}_{i-1})) \le \sum_{i=1}^{n} H(\mathbf{x}_i) \quad (4)$$

Assume that these variables are disentangled into three factors tumor, microenvironment, and background. Let bag $\mathbf{X}$ be partitioned into three subsets with sizes n_1, $n_2 - n_1$ and $n - n_2$. We can derive that:

$$H^*(\mathbf{x}_1, \cdots, \mathbf{x}_n) = w_1 \sum_{i=1}^{n_1} H(\mathbf{x}_i|\boldsymbol{\Phi}^{-1}(\mathbf{x}_i) \in \mathbf{h}_1 = \mathbf{T}) + w_2 \sum_{i=n_1+1}^{n_2} H(\mathbf{x}_i|\boldsymbol{\Phi}^{-1}(\mathbf{x}_i) \in \mathbf{h}_2 = \mathbf{E})$$

$$+ w_3 \sum_{i=n_2+1}^{n} H(\mathbf{x}_i|\boldsymbol{\Phi}^{-1}(\mathbf{x}_i) \in \mathbf{h}_3 = \epsilon) \le H(\mathbf{x}_1) + \sum_{i=2}^{n} H(\mathbf{x}_i|\boldsymbol{\Phi}^{-1}(\mathbf{x}_1, \cdots, \mathbf{x}_{i-1}))$$

$$\le \sum_{i=1}^{n} H(\mathbf{x}_i)$$

$$(5)$$

where H^* denotes the correlated information entropy and w_j denotes the instance effects. The weighted summation after semantic disentanglement only decrease the entropy value compared to the original fuzzy decision. Therefore, the total information entropy of a bag in (5) decreases, which enhance the ability to mitigate decision entanglement.

2   Algorithm

In Algorithm 1, we present the detailed process of Cluster-reasoning Instance Disentangling (CID) and Instance Effect Re-weighting in PG-CIDL

Algorithm 1 CID and Instance Effect Re-weighting in PG-CIDL

Input: Instance groups $\mathbf{Z} = \{\mathbf{Z}_1, \mathbf{Z}_2, \mathbf{Z}_3\}$; each $\mathbf{Z}_k$ belongs to one of the factors TC, ME and BG.

Output: Final bag feature $\mathbf{Z}_{\text{final}}$ and factor map `factor`

1: $P_v \leftarrow f\big(\sigma(\mathbf{Z})\big)$ ▷ vanilla prediction
2: **for** $k = 1$ to 3 **do**
3:      $\mathbf{Z}_{\backslash k} \leftarrow \big\{\mathbf{Z}_j : j \neq k\big\}$
4:      $P_k \leftarrow f\big(\sigma(\mathbf{Z}_{\backslash k})\big)$
5:      $D_k \leftarrow D_{\text{KL}}\big(P_v \| P_k\big)$
6: **end for**
7: $\{w_k\}_{k=1}^3 \leftarrow$ `Normalize`$(\{D_k\})$
8: `order` $\leftarrow$ `argsort`$(\{w_k\}, \text{descend})$
9: `factor`[TC] $\leftarrow \mathbf{Z}_{\text{order}[1]}, \dots,$ `factor`[BG] $\leftarrow \mathbf{Z}_{\text{factor}[3]}$
10: $\mathbf{Z}_{\text{final}} \leftarrow$ `mean`$\big(\sum_{k=1}^3 w_k \mathbf{Z}_k\big)$
11: **return** $\mathbf{Z}_{\text{final}}$, `factor`

---

3 Datasets and Setup

3.1 Datasets

To validate PG-CIDL for diagnosis and sub-typing tasks, we conduct extensive experiments on two public datasets and two private datasets. Two private WSI datasets were collected from some institution, where **AMU-LSCC** is for laryngeal squamous cell carcinoma pathological grading and **AMU-CSCC** is for cervical squamous cell carcinom. 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. Two chosen public datasets are **CAMELYON16** for breast cancer lymph node metastasis detection [2] and **DHMC-LUNG** [6] for lung adenocarcinoma classification.

3.2 Implementation

For comprehensive comparison, we extract features with Swin Transformer [3] pretrained on ImageNet-11k. We utilize Rmsprop optimizer for training. Models on two private datasets were trained with a batch size of 2 and 100 epochs. For public datasets, batch size is set 1. The random seed was fixed 0 across all stages. The input resolution of PG-CIDL is $96 \times 96 \times 3$. The learning rate schedule was: 1×10^{-5} for epochs 1–50, 5×10^{-6} for epochs 51–75, and 1×10^{-6} for epochs 76–100. The hyper-parameter γ in loss function was set 0.1. Patches were grid cropped from each WSI in 10x magnification and we filtered out instances with blank areas.

3.3   Computational Specs

Experiments were conducted on Ubuntu 22.04 with x86_64 architecture with four *NVIDIA A10 Tensor Core 24GB*. For computational framework and library versions: we use PyTorch 2.6.0, CUDA 12.4, cuDNN 9.1. For training CAMELYON16 specifically, the peak memory usage is 13GB each GPU at one batch. In terms of model inference, it takes 0.123s per WSI for feature extraction and 0.355s for three-phase DRL.

4   Experimental Results

4.1   ROC plots

Figure 2 illustrates the curve of PG-CIDL lies above all the other curves and suggests the largest area under the curve (AUC), which implies that our PG-CIDL achieves the highest predictive confidence among all SOTA models.
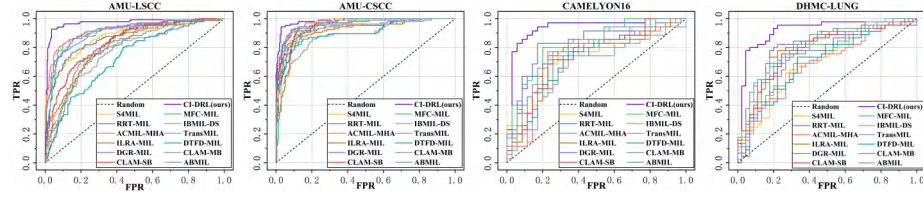

Fig. 2. ROC plots' comparison on multicentre datasets.

4.2   Extra visualization results.

Here, we provide extra visualization results of our PG-CIDL on AMU-LSCC and DMHC-LUNG; see Fig. 3.

4.3   Capability of Feature Representation

We utilize UMAP [1] to visualize feature representation in 2-D space. In Figure 4, we can see that our proposed PG-CIDL show less entanglement within feature distributions, while other sub-optimal models without spatial disentanglement show more feature overlaps. It not only implies that PG-CIDL disentangles feature spatially, but it also has superior ability to distinguish feature related to corresponding grading results.

An effective classification model should be able to capture and fit the differences among pathological patterns. To validate this assumption, we visualize the final layer feature representation of each model using UMAP in 1-D space. Figure 5 shows the analysis of variance (ANOVA) with effect size metrics (η^2).
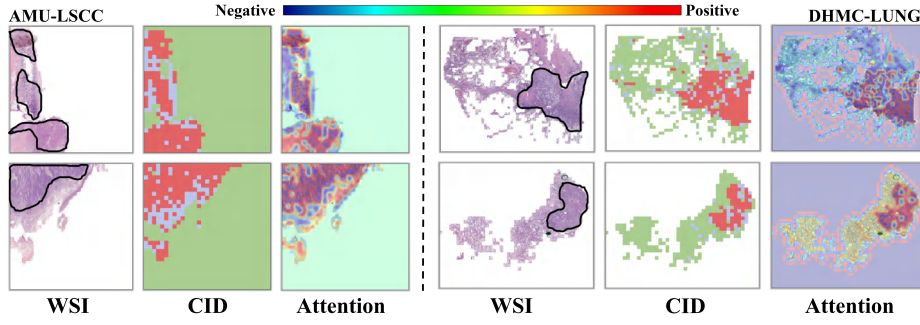
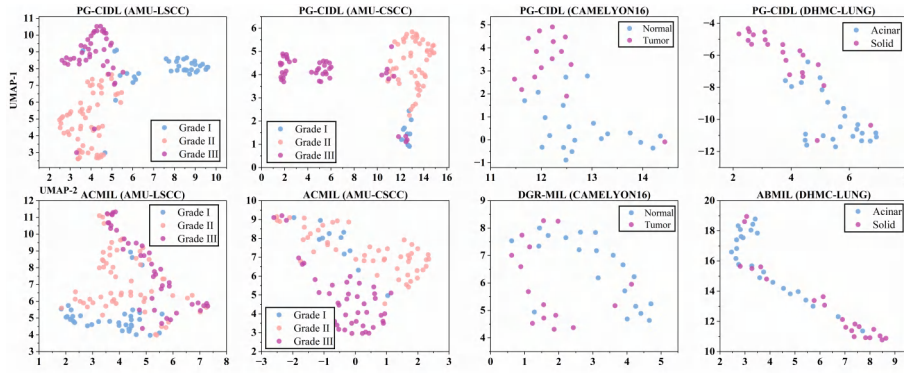**Fig. 3.** Extra visualization results on AMU-LSCC and DHMC-LUNG.

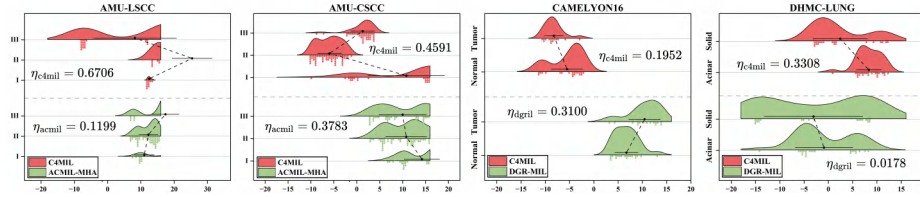**Fig. 4.** 2-D feature representation with spatial disentanglement on multicentre datasets.

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

4.4 Model Complexity Analysis

End-to-end learning for WSI analysis is considered to be computationally intensive, accordingly receiving limited exploration. However, our findings suggest that its actual computational demand has been overestimated. In Supplemen-
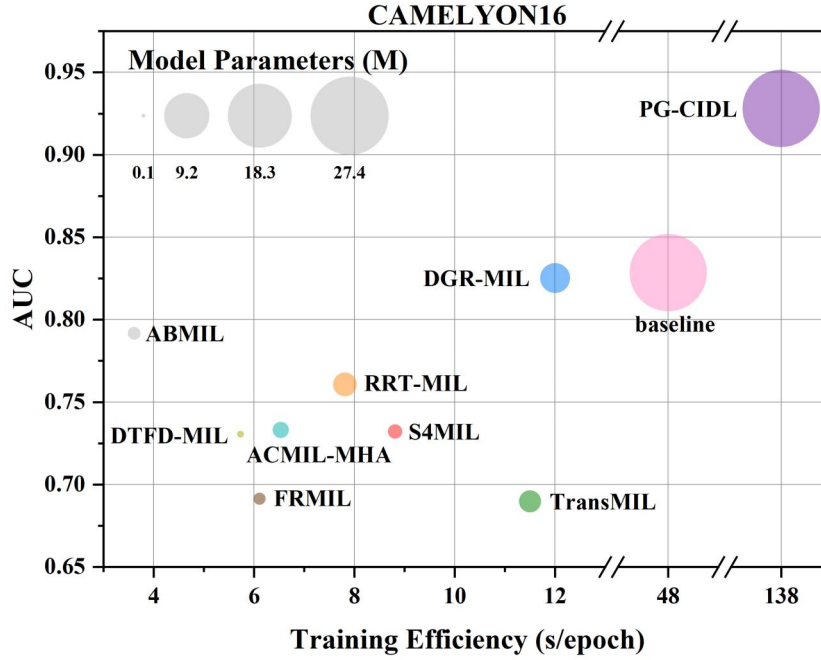
Fig. 6. Model complexity vs. performance.

tary Material, we present detailed computational specs. In Figure 6 we present comparison of complexity vs. performance on CAMELYON16 of batch size 1.

Specifically, our PG-CIDL contains 27.50M parameters significantly more than existing models, which range from 0.199M (DTFD-MIL) to 4.075M (DGR-MIL). This leads to a longer training time per epoch for PG-CIDL (138s), compared to 12.0s for DGR-MIL, the most computationally demanding among the SOTA models. Although our PG-CIDL approach increases training time ten times larger, the computational cost is relatively accepted. It allows the model to learn highly task-specific feature representations, resulting in better classification accuracy and enhanced interpretability.

References

1. Becht, E., McInnes, L., Healy, J., Dutertre, C.A., Kwok, I.W., Ng, L.G., Ginhoux, F., Newell, E.W.: Dimensionality reduction for visualizing single-cell data using umap. Nature biotechnology **37**(1), 38–44 (2019)
2. Bejnordi, B.E., Veta, M., Van Diest, P.J., Van Ginneken, B., Karssemeijer, N., Litjens, G., Van Der Laak, J.A., Hermsen, M., Manson, Q.F., Balkenhol, M., 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)

3. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
4. 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)
5. Wang, X., Chen, H., Tang, S., Wu, Z., Zhu, W.: Disentangled representation learning. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(12), 9677–9696 (2024)
6. Wei, J.W., Tafe, L.J., Linnik, Y.A., Vaickus, L.J., Tomita, N., Hassanpour, S.: Pathologist-level classification of histologic patterns on resected lung adenocarcinoma slides with deep neural networks. Scientific reports **9**(1), 3358 (2019)