Cracking Instance Jigsaw Puzzles: An Alternative to Multiple Instance Learning for Whole Slide Image Analysis

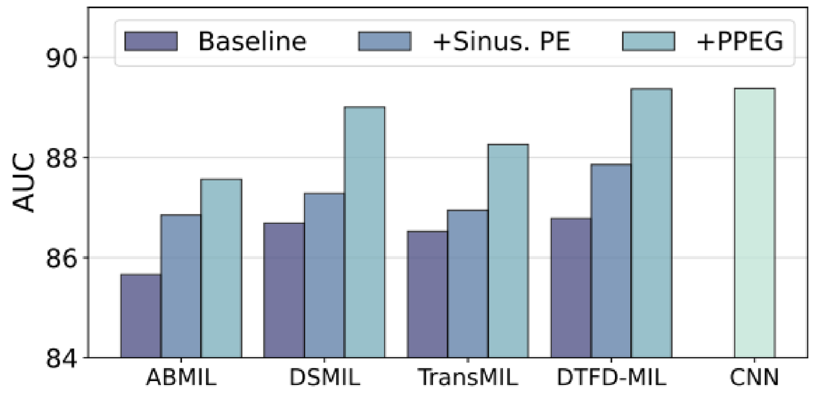
Xiwen Chen[1]* Peijie Qiu[2]* Wenhui Zhu[3]* Hao Wang[1] Huayu Li[4] Xuanzhao Dong[3]
Xiaotong Sun[5] Xiaobing Yu[2] Yalin Wang[3] Abolfazl Razi[1] Aristeidis Sotiras[2]
1. Clemson University 2. Washington University in St. Louis 3. Arizona State University
4. University of Arizona 5. University of Arkansas
* equal contributions

Paper

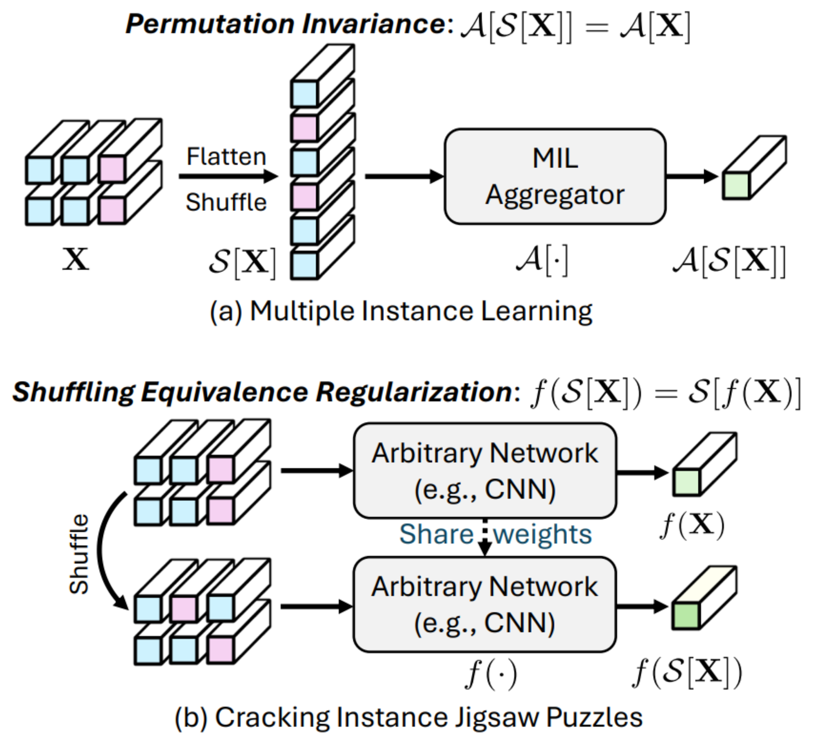
ICCV HONOLULU HAWAII OCT 19-23, 2025

Motivation

- **Permutation invariance** is a core assumption in MIL but it overlooks spatial correlations between tiles in WSIs, which are essential for image-based analysis. **Neighboring tiles** in WSIs are usually spatially and semantically related, often belonging to the same tissue category.
- This creates a **dilemma** between maintaining permutation invariance and preserving spatial correlations.
- Convolutional operations are effective for modeling spatial structure but they are not permutation-invariant because rearranging the inputs changes the output. **Adding convolutional layers** to MIL aggregators often leads to clear performance improvements. **Replacing traditional MIL aggregators with simple CNNs** can even achieve better performance, a direction that has received little attention in the literature.
- These facts lead us to question whether we really need an MIL for WSI analysis. Instead, we argue that, *despite violating the permutation-invariant constraint, modeling the spatial correlations between tiles/instances is necessary*.



Methods



(a) the traditional MIL method and **(b)** the proposed method for solving instance jigsaw puzzles. Compared to MIL, the proposed method is advantageous in uncovering semantic correlations between instances.

Methods

Rather than achieving better modeling instance correlation through different attention mechanisms, naive position encoding e.g., sin-cos, PPEG [1]), or accessing additional information, we discover **that learning to restore the original order from shuffled tiles can lead to better performance by exploiting their semantic correlations.** We term this process as solving an **instance jigsaw puzzle**.

- **A Siamese Network Solution**

To solve the above instance jigsaw puzzles problem, we propose a Siamese network solution. Specifically, given a shuffling operator $\mathcal{S}[\cdot]$, which randomly shuffles input instances, the following equivalence should hold for solving the instance jigsaw puzzles:

$$f_\theta(\mathcal{S}[\mathbf{X}]) = \mathcal{S}[f_\theta(\mathbf{X})], \tag{2}$$

where f_θ is parameterized by a neural network, the rationale is that if a network can learn to restore the correct arrangement from shuffled instances, applying the inverse shuffling operation $\mathcal{S}^{-1}[\cdot]$ to the network's output should recover the original arrangement: $\mathcal{S}^{-1}[f_\theta(\mathcal{S}[\mathbf{X}])] = f_\theta(\mathbf{X})$. Accordingly, we design a shuffling equivalence regularization loss as follows:

$$\mathcal{L}_{\text{Equv}}(\mathbf{X}) = \frac{1}{2n}\|f(\mathcal{S}[\mathbf{X}]) - \mathcal{S}[f(\mathbf{X})]\|_2^2, \tag{3}$$

which penalizing the mean squared error between $f(\mathcal{S}[\mathbf{X}])$ and $\mathcal{S}[f(\mathbf{X})]$. This equivalence loss is then implemented as a Siamese network with two branches that share the same weights (see Fig. 1(b)). The first branch takes as input the unshuffled instances $\mathbf{X}$, while the second branch takes as input the shuffled instances $\mathcal{S}[\mathbf{X}]$. The final objective for WSI classification is then a weighted combination of the equivalence loss and the binary cross-entropy loss ($\mathcal{L}_{\text{BCE}}$):

$$\mathcal{L}_{\text{final}}(\mathbf{X}, \mathbf{Y}) = \mathcal{L}_{\text{BCE}}(\mathbf{X}, \mathbf{Y}) + \lambda\mathcal{L}_{\text{Equv}}(\mathbf{X}), \tag{4}$$

Theorem 3. *When approximating the optimal transport plan $\mathbf{T}_\#$ with the inverse shuffling operation $\mathcal{S}^{-1}$, the proposed shuffling equivalence regularization is the solution to the inverse optimal transport problem.*

Experiment and Ablation

Table 1. Main results on the CAMELYON16 dataset and TCGA-NSCLC dataset by using different feature extractors. Our method significantly outperforms all MIL-based competitors (see Appendix E for the statistical test).

	CAMELYON16			TCGA-NSCLC		
	Accuracy	F1	AUC	Accuracy	F1	AUC
Swin-VIT ImageNet Pretrained						
ABMIL (*ICML'18*)	$84.73_{0.85}$	$83.20_{0.81}$	$85.66_{1.76}$	$91.07_{1.08}$	$91.27_{1.23}$	$95.88_{1.18}$
DSMIL (*CVPR'21*)	$84.42_{1.12}$	$82.72_{1.15}$	$86.69_{2.33}$	$90.98_{1.49}$	$90.97_{1.49}$	$95.71_{0.18}$
TransMIL (*NeurIPS'21*)	$85.04_{1.70}$	$83.72_{1.29}$	$88.26_{0.88}$	$89.73_{0.40}$	$89.93_{0.62}$	$95.66_{0.99}$
MaxS (*CVPR'22*)	$84.57_{1.22}$	$78.87_{1.13}$	$89.69_{1.25}$	$87.33_{1.00}$	$87.05_{1.31}$	$93.09_{0.85}$
AFS (*CVPR'22*)	$79.61_{2.22}$	$72.18_{0.95}$	$83.88_{1.68}$	$90.79_{1.52}$	$90.36_{1.80}$	$96.17_{0.89}$
MaxMinS (*CVPR'22*)	$83.80_{1.01}$	$76.73_{1.29}$	$86.78_{1.59}$	$89.83_{0.87}$	$89.44_{1.22}$	$95.76_{0.57}$
ILRA-MIL (*ICLR'23*)	$84.96_{1.05}$	$83.60_{0.86}$	$87.76_{1.45}$	$90.69_{1.13}$	$90.68_{1.13}$	$95.56_{0.97}$
MHIM-MIL (*ICCV'23*)	$86.24_{1.68}$	$84.35_{2.15}$	$86.12_{1.95}$	$89.64_{1.66}$	$89.61_{1.67}$	$93.93_{0.84}$
DGR-MIL (*ECCV'24*)	$87.60_{2.39}$	$86.47_{2.39}$	$88.19_{1.73}$	$90.88_{1.83}$	$90.85_{1.84}$	$95.81_{1.25}$
AC-MIL (*ECCV'24*)	$86.24_{1.01}$	$84.94_{1.31}$	$87.77_{1.61}$	$90.50_{1.29}$	$90.63_{1.23}$	$95.61_{0.79}$
Ours [Trans.]	**$89.53_{1.40}$**	**$88.57_{1.53}$**	**$92.17_{0.49}$**	*$92.32_{1.25}$*	*$92.31_{1.26}$*	**$96.40_{0.77}$**
Ours [CNN]	*$88.11_{0.58}$*	*$87.11_{0.63}$*	*$91.80_{0.26}$*	**$92.51_{0.80}$**	**$92.49_{0.81}$**	*$96.32_{0.67}$*
ResNet-18 ImageNet Pretrained						
ABMIL (ICML'18)	$85.74_{0.99}$	$84.21_{1.11}$	$85.91_{1.53}$	$88.10_{0.80}$	$88.18_{0.82}$	$93.88_{1.11}$
DSMIL (*CVPR'21*)	$84.19_{2.25}$	$82.21_{2.82}$	$84.84_{1.74}$	$88.58_{1.02}$	$88.61_{1.06}$	$93.73_{0.87}$
TransMIL (*NeurIPS'21*)	$82.79_{1.89}$	$76.63_{1.86}$	$87.71_{1.84}$	$84.65_{1.11}$	$84.20_{0.90}$	$90.71_{1.20}$
MaxS (*CVPR'22*)	$84.81_{2.09}$	$83.62_{2.20}$	$87.22_{1.78}$	$88.39_{0.81}$	$88.54_{1.00}$	$93.43_{0.84}$
AFS (*CVPR'22*)	$81.94_{1.55}$	$77.85_{1.45}$	$89.23_{1.07}$	$88.48_{0.81}$	$88.27_{1.16}$	$94.83_{0.92}$
MaxMinS (*CVPR'22*)	$82.02_{1.86}$	$76.11_{0.88}$	$88.04_{1.84}$	$87.81_{0.86}$	$87.5_{1.00}$	$94.19_{0.95}$
ILRA-MIL (*ICLR'23*)	$87.08_{2.31}$	$86.19_{2.56}$	$89.30_{2.98}$	$88.77_{0.98}$	$88.8_{1.09}$	$94.25_{0.68}$
MHIM-MIL (*ICCV'23*)	$86.05_{1.64}$	$84.48_{1.82}$	$86.17_{1.76}$	$87.43_{1.37}$	$87.4_{1.35}$	$93.65_{0.62}$
DGR-MIL (*ECCV'24*)	$86.63_{0.85}$	$85.25_{0.96}$	$88.20_{1.30}$	$87.43_{1.18}$	$87.43_{1.14}$	$93.88_{0.41}$
AC-MIL (*ECCV'24*)	$87.02_{1.49}$	$85.55_{1.77}$	$87.56_{2.37}$	$88.58_{0.69}$	$88.58_{0.69}$	$94.31_{1.12}$
Ours [Trans.]	*$87.47_{2.12}$*	*$86.30_{2.34}$*	*$90.44_{1.41}$*	*$88.96_{0.97}$*	*$89.02_{0.98}$*	**$94.98_{0.81}$**
Ours [CNN]	**$88.37_{0.45}$**	**$87.16_{0.36}$**	**$92.92_{0.87}$**	**$90.40_{0.98}$**	**$90.39_{0.98}$**	*$94.93_{1.15}$*
CTransPath Self-supervised Pretrained						
ABMIL (*ICML'18*)	$94.80_{0.50}$	$94.39_{0.55}$	$96.50_{0.67}$	$93.38_{1.10}$	$93.36_{1.11}$	$96.81_{0.63}$
DSMIL (*CVPR'21*)	$94.49_{0.64}$	$94.08_{0.69}$	$95.64_{0.56}$	$94.24_{1.25}$	$94.22_{1.26}$	$97.85_{0.69}$
TransMIL (*NeurIPS'21*)	$94.42_{0.58}$	$92.44_{0.68}$	$97.34_{0.19}$	$90.79_{0.72}$	$90.39_{0.69}$	$96.22_{0.79}$
MaxS (*CVPR'22*)	$94.96_{1.21}$	$94.77_{1.15}$	$97.33_{0.30}$	$93.86_{0.95}$	$93.85_{0.94}$	$97.84_{0.52}$
AFS (*CVPR'22*)	$94.42_{0.68}$	$92.42_{0.86}$	$97.14_{0.27}$	$93.28_{1.04}$	$92.95_{1.17}$	$97.81_{0.46}$
MaxMinS (*CVPR'22*)	$95.19_{0.47}$	$93.38_{0.54}$	$97.66_{0.44}$	$93.66_{0.70}$	$93.34_{0.74}$	$97.78_{0.39}$
ILRA-MIL (*ICLR'23*)	$94.83_{1.77}$	$94.45_{1.94}$	$95.85_{1.00}$	$93.57_{0.75}$	$93.56_{0.75}$	$97.44_{0.56}$
MHIM-MIL (*ICCV'23*)	$94.57_{0.55}$	$94.16_{0.65}$	$96.38_{0.61}$	$93.95_{1.21}$	$93.94_{1.21}$	*$97.87_{0.53}$*
DGR-MIL (*ECCV'24*)	$95.73_{1.16}$	$95.41_{1.26}$	$96.30_{0.47}$	$94.53_{1.26}$	$94.52_{1.27}$	*$97.87_{0.50}$*
AC-MIL (*ECCV'24*)	$95.15_{0.64}$	$94.88_{0.73}$	$97.00_{0.69}$	$94.72_{0.68}$	$94.72_{0.68}$	$97.76_{0.76}$
Ours [Trans.]	**$96.64_{0.37}$**	**$96.39_{0.38}$**	*$98.00_{0.17}$*	**$95.20_{1.23}$**	**$95.19_{1.23}$**	**$97.99_{0.67}$**
Ours [CNN]	*$96.25_{0.83}$*	*$95.99_{0.88}$*	**$98.10_{0.31}$**	*$95.11_{0.83}$*	*$95.10_{0.84}$*	$97.55_{0.77}$

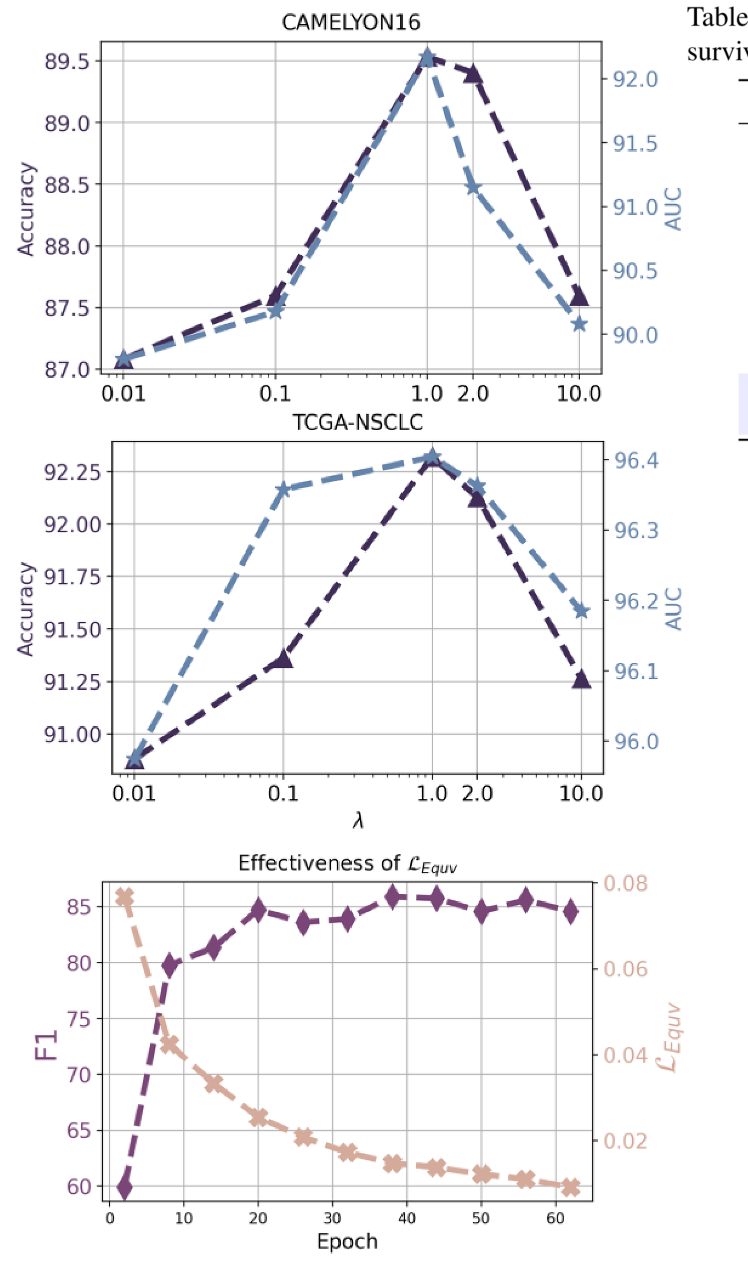


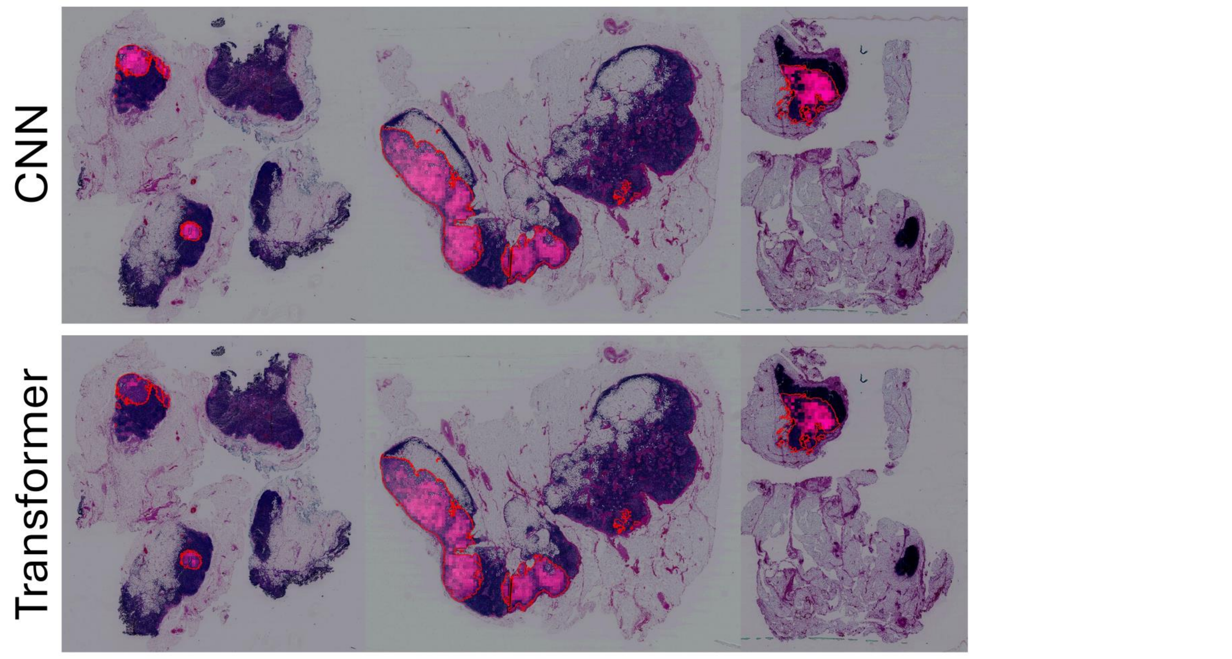
Figure 3. **Top:** The effectiveness of λ on CAMELYON16 datasets, **Middle:** the effectiveness of λ on TCGA-NSCLC datasets, and **Bottom:** Training dynamics for $\mathcal{L}_{\text{Equv}}$ and performance.

Table 2. Comparison of the performance by different methods for survival prediction in terms of *C-Index* using **UNI** [9] features.

	TCGA-LUAD	TCGA-BRCA
ABMIL *(ICML'18)*	$0.659_{0.022}$	$0.728_{0.050}$
DSMIL *(CVPR'21)*	$0.656_{0.029}$	$0.715_{0.052}$
TransMIL *(NeurIPS'21)*	$0.609_{0.032}$	$0.721_{0.023}$
DTFD-MIL *(CVPR'22)*	$0.656_{0.046}$	$0.733_{0.022}$
ILRA-MIL *(ICLR'23)*	$0.595_{0.032}$	$0.718_{0.026}$
MHIM-MIL *(ICCV'23)*	$0.650_{0.020}$	$0.736_{0.031}$
DGR-MIL *(ECCV'24)*	$0.620_{0.013}$	$0.711_{0.038}$
AC-MIL *(ECCV'24)*	$0.656_{0.017}$	$0.659_{0.062}$
Ours [Trans.]	*$0.677_{0.028}$*	*$0.752_{0.017}$*
Ours [CNN]	**$0.683_{0.030}$**	**$0.757_{0.018}$**

Table 3. The ablation studies on the shuffling equivalence regularization ($\mathcal{L}_{\text{Equv}}$) on CAMELYON16 and TCGA-NSCLC. Please refer to Appendix B.4 for ablation using **CtransPath** features.

Dataset	Network	$\mathcal{L}_{\text{Equv}}$	Accuracy	F1	AUC
CAMELYON16	**Trans.**	✗	86.51	85.17	88.89
		✓	89.53	88.57	92.17
		Δ	**+3.02**	**+3.40**	**+3.28**
	CNN	✗	85.58	84.12	89.38
		✓	88.11	87.11	91.80
		Δ	**+2.53**	**+2.99**	**+2.43**
TCGA-NSCLC	**Trans.**	✗	90.60	90.92	95.80
		✓	92.32	92.31	96.40
		Δ	**+1.72**	**+1.39**	**+0.60**
	CNN	✗	90.21	90.20	95.65
		✓	92.51	92.49	96.32
		Δ	**+2.30**	**+2.29**	**+0.67**



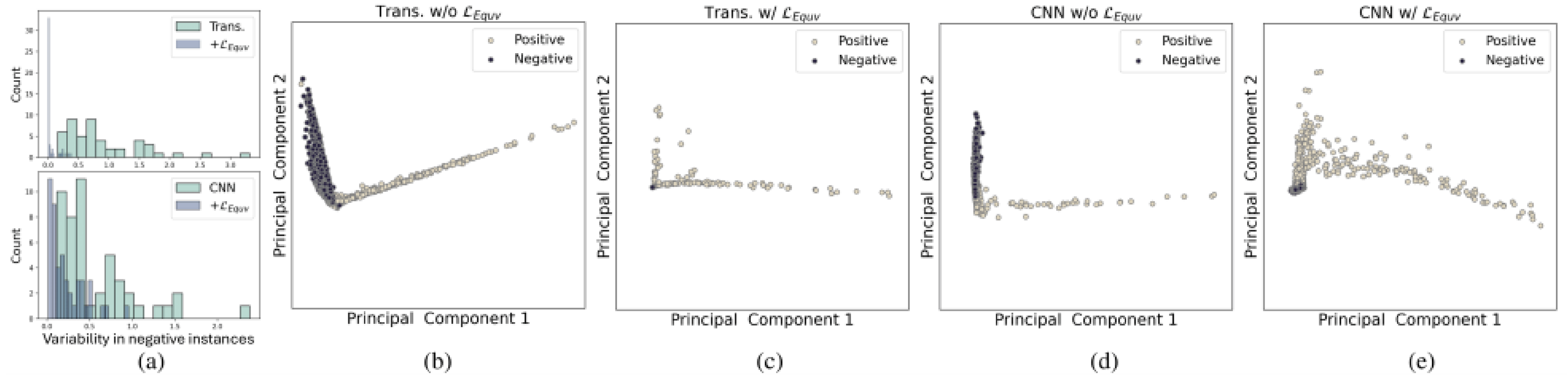


Figure 4. (a) The histogram of variability of the negative instance feature representations from different WSIs on CAMELYON16 (please refer to [29, Eq. 4] for details). (b-e) The learned instance-level feature representations by different variants of our method are visualized by projecting them into two dimensions using PCA. The proposed method with equivalence regularization shows small variability, which corresponds to the collapse of variability when the learned representations converge to a compact and uniform structure without redundancy.

Conclusion

We question the necessity of traditional permutation-invariant MIL for whole slide image analysis and highlight the critical role of spatial and semantic correlations among instances.To address this, we propose **"Cracking Instance Jigsaw Puzzles"**, a new paradigm that explicitly learns semantic relationships by restoring the spatial order of shuffled tiles. Our **Siamese network with shuffling equivalence regularization**, theoretically supported by optimal transport principles, consistently outperforms state-of-the-art MIL models in both **WSI classification** and **survival prediction** tasks. Despite a minor computational cost, the approach broadens the design space by enabling flexible architectures such as CNNs, paving the way for more spatially aware and semantically enriched WSI analysis.

Reference

[1] Shao, Zhuchen, et al. "Transmil: Transformer based correlated multiple instance learning for whole slide image classification." Advances in neural information processing systems 34 (2021): 2136-2147.