Rethinking Pre-Trained Feature Extractor Selection in Multiple Instance Learning for Whole Slide Image Classification

Bryan Wong[1], Sungrae Hong[1], Mun Yong Yi[2]

[1]Graduate School of Data Science, KAIST, Daejeon, South Korea
[2]Department of Industrial and Systems Engineering, KAIST, Daejeon, South Korea

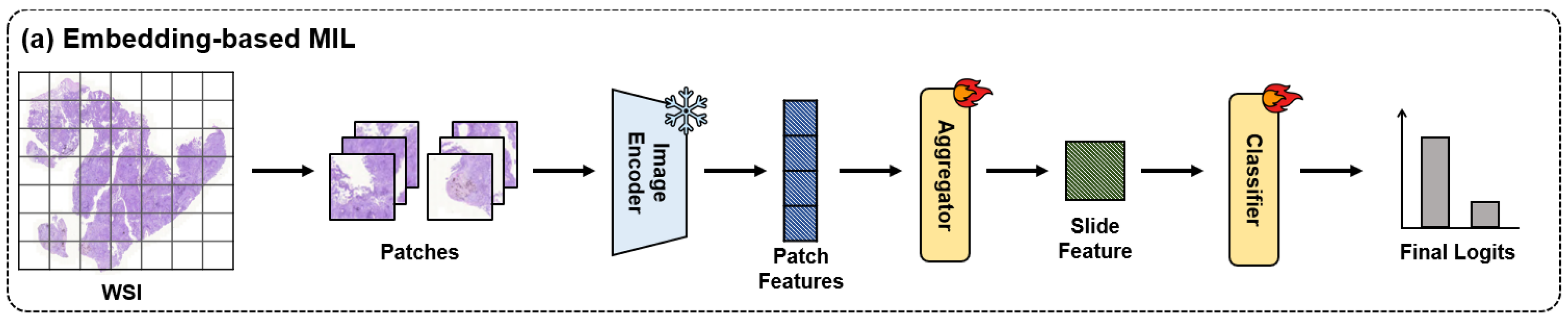
ISBI 2025
2025 IEEE International Symposium on Biomedical Imaging
April 14-17, 2025 | Houston, TX, USA

Poster No. Session 2 - 63 (1571090653)

INTRODUCTION

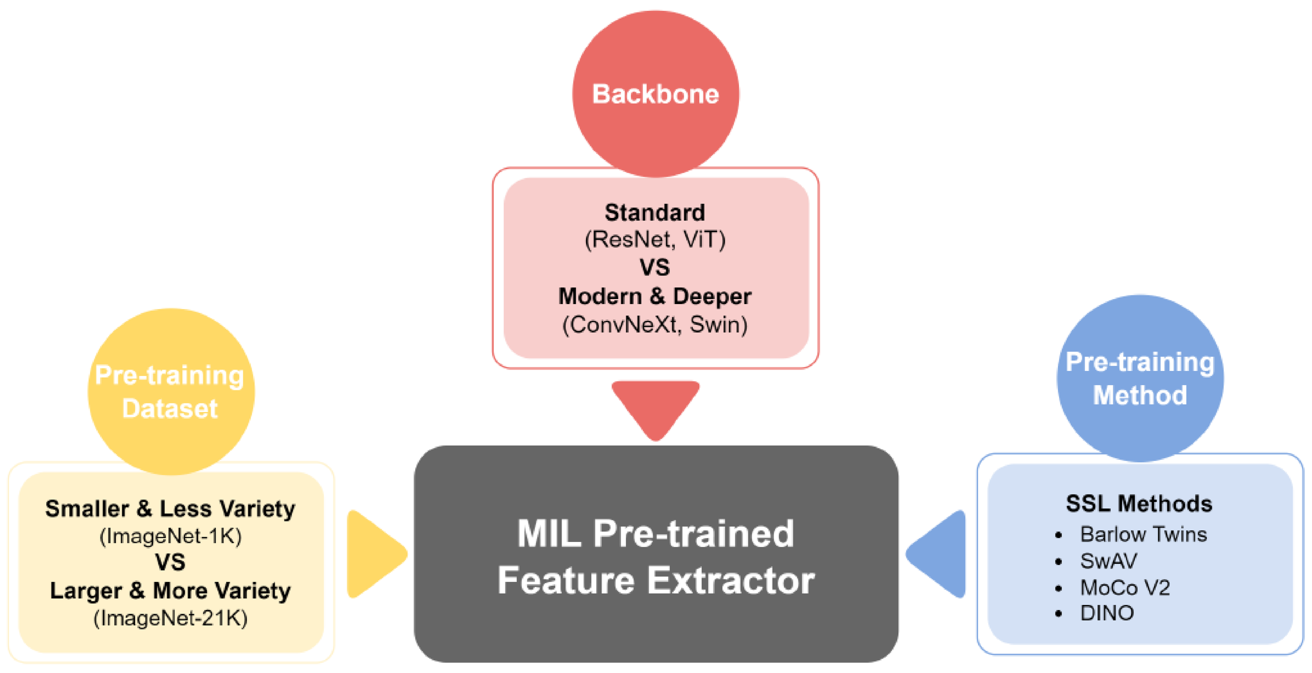
✓ Background

- **Multiple instance learning (MIL)** has become a preferred method for gigapixel whole slide image (WSI) classification **without requiring patch-level annotations**
- Current research primarily relies on **embedding-based MIL** approaches, which extract patch features using a pre-trained feature extractor and aggregate them for slide-level prediction



- Despite the critical role of feature extraction, *there is limited guidance on selecting optimal feature extractors to maximize WSI performance*
- This study addresses this gap by systematically **evaluating MIL feature extractors across three dimensions: pre-training dataset, backbone model, and pre-training method**
- Using **two public WSI datasets** (TCGA-NSCLC and Camelyon16) and employing **four MIL models** (ABMIL, DSMIL, TransMIL, and DTFD-MIL), this **study is the first to undertake a comprehensive analysis focused on optimal feature extractor selection**

ANALYSIS SETUP



1. Pre-training Dataset

- Most MIL models use feature extractors pre-trained on ImageNet-1K
- Recent studies show ImageNet-21K improves transferability and performance
- We explore whether **larger, more diverse pre-training datasets lead to better WSI classification in MIL**

2. Backbone

- Most MIL models use standard backbones like ResNet (CNN) and ViT (Transformer)
- We evaluate whether **modern, larger backbones** (e.g., ConvNeXt-B, Swin-B) —pre-trained with the same dataset and method—**can generate stronger features that improve MIL robustness and generalization**

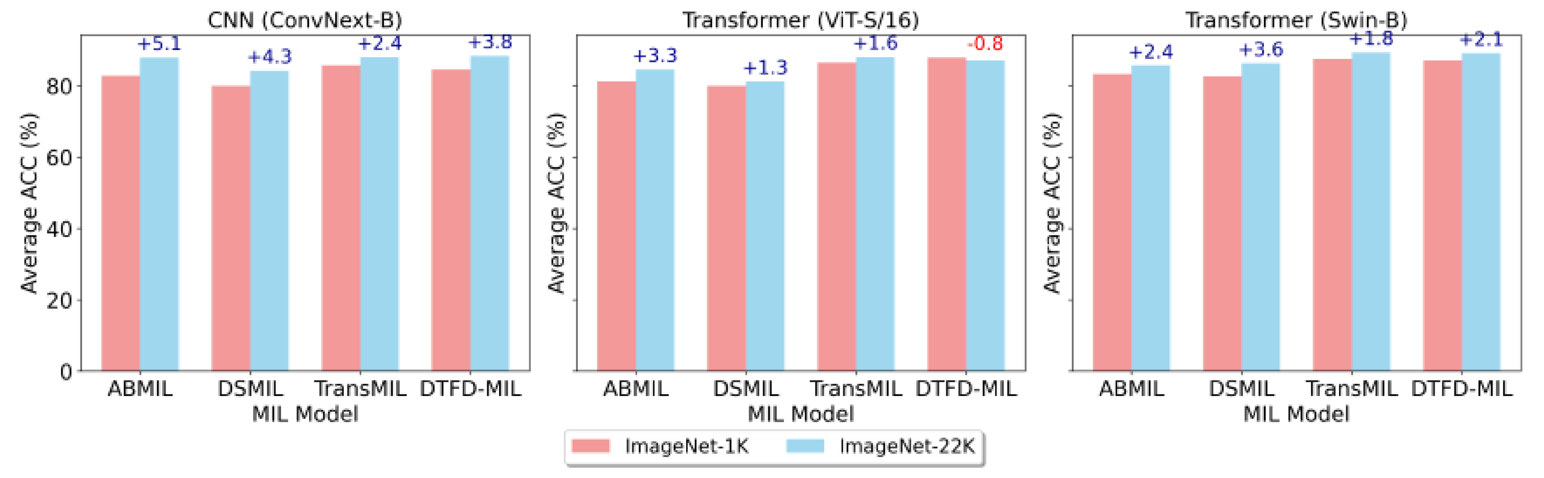
3. Pre-training Method

- Self-supervised learning (SSL) is well-suited for medical imaging, where labeled data are limited
- Yet, it's **unclear which SSL method best enhances MIL performance**
- We compare **four representative SSL approaches**: **contrastive** (MoCoV2), **non-contrastive** (Barlow Twins), **clustering** (SwAV), and **ViT-based SSL** (DINO)

ACKNOWLEDGEMENTS: This research was supported by the Seegene Medical Foundation, South Korea, under the project "Development of a Multimodal Artificial Intelligence-Based Computer-Aided Diagnosis System for Gastrointestinal Endoscopic Biopsies" (Grant Number: G01240151).

EXPERIMENT

✓ Pre-training Dataset Size and Variety



- **Pre-training on ImageNet-21K consistently improves** WSI classification compared to ImageNet-1K across **both CNN** (ConvNext-B) and **Transformer** (ViT-S/16, Swin-B) backbones
- The performance gain is **backbone-independent**, suggesting that larger and more diverse pre-training datasets yield richer feature representations and stronger generalization

✓ Standard vs Modern Backbones

MIL Model	TCGA-NSCLC ResNet50 8.5M, IN-1K ACC	AUC	ConvNeXt-B 43.2M, IN-21K ACC	AUC	ViT-S/16 21.7M, IN-21K ACC	AUC	Swin-B 86.7M, IN-21K ACC	AUC	Camelyon16 ResNet50 8.5M, IN-1K ACC	AUC	ConvNeXt-B 43.2M, IN-21K ACC	AUC	ViT-S/16 21.7M, IN-21K ACC	AUC	Swin-B 86.7M, IN-21K ACC	AUC
ABMIL	85.08	89.90	**87.94**	92.97	84.60	**93.13**	85.72	90.12	78.30	74.84	75.64	61.79	78.04	72.61	**81.65**	**77.62**
DSMIL	84.29	91.24	84.28	92.70	81.27	90.32	**86.35**	**93.60**	**80.36**	**77.57**	79.33	75.01	73.64	75.48	74.68	73.11
TransMIL	85.08	90.96	88.09	93.44	88.10	**93.95**	**89.37**	89.18	80.62	80.77	78.55	78.86	79.59	78.58	**88.11**	**89.86**
DTFD-MIL	87.78	94.34	88.41	94.12	87.14	93.07	**89.21**	**94.52**	82.17	86.53	80.28	85.12	80.62	81.89	**88.11**	**90.47**
Average	85.56	91.61	87.18	**93.31**	85.28	92.62	**87.66**	91.85	80.36	79.93	**78.45**	**75.20**	77.97	77.14	**83.14**	**82.77**

- **Modern and deeper backbones** (ConvNeXt-B, Swin-B) **outperform standard ones** (ResNet50, ViT-S/16) on TCGA-NSCLC
- On **Camelyon16, the modern Transformer-based (Swin-B) outperforms all others**, while the modern CNN-based (ConvNeXt-B) underperforms compared to traditional CNN-based (ResNet50)
- This highlights **Transformers' advantage in modeling fine-grained small patterns** (e.g., small tumor ROIs) via self-attention and their ability to **scale reliably with deeper architecture**

✓ Self-Supervised Pre-training Methods

MIL Model	TCGA-NSCLC Barlow Twins ResNet50 ACC	AUC	SwAV ResNet50 ACC	AUC	MoCo V2 ResNet50 ACC	AUC	DINO ViT-S/16 ACC	AUC	Camelyon16 Barlow Twins ResNet50 ACC	AUC	SwAV ResNet50 ACC	AUC	MoCo V2 ResNet50 ACC	AUC	DINO ViT-S/16 ACC	AUC
ABMIL	87.78	93.99	85.87	93.27	85.71	89.86	**90.48**	**96.83**	91.47	92.06	**94.83**	**95.14**	76.74	73.03	94.06	94.57
DSMIL	86.35	93.97	85.72	93.53	76.03	89.08	**89.05**	**96.34**	88.63	87.88	92.25	91.26	65.38	66.95	**95.09**	**98.25**
TransMIL	89.68	92.11	89.84	**95.69**	88.41	92.55	**92.86**	95.64	93.28	94.26	94.83	96.64	93.02	94.4	**97.15**	**98.10**
DTFD-MIL	89.21	91.97	89.52	95.66	70.32	76.82	**93.18**	**97.62**	91.18	94.97	94.83	96.46	64.60	63.16	**97.41**	**98.07**
Average	88.26	93.01	87.74	94.54	**80.12**	**87.08**	**91.39**	**96.61**	91.14	92.29	94.19	94.92	**74.94**	**74.37**	**95.93**	**97.25**

- **SSL method and backbone choice** greatly affect WSI classification
- **DINO + ViT-S/16 outperforms others**; **MoCoV2 + ResNet50 performs worst on Camelyon16** due to contrastive loss sensitivity due to tissue similarity and class imbalance
- **Pre-training method matters more than dataset domain**—poor SSL (e.g., MoCoV2) on in-domain data can underperform compared to ImageNet pre-training

CONCLUSION

- Results show that **SSL method choice has greater impact than in-domain dataset** selection alone
- We recommend **Transformer-based backbones with deeper architectures** over CNNs for improved generalization
- We also recommend **larger, more diverse pre-training datasets** to enhance feature quality and downstream performance

RESOURCES

Paper

Code

KAIST GSDS
Graduate School of Data Science

KAIST ISE
Department of Industrial and Systems Engineering

Seegene

KNOWLEDGE INNOVATION RESEARCH CENTER