LRMIL: Efficient Low-Resolution Multiple Instance Learning via High-Resolution Knowledge Distillation for Whole Slide Image Classification

Yonghan Shin, Won-Ki Jeong†

Department of Computer Science and Engineering, Korea University, Seoul, Korea
{dydgks592, wkjeong}@korea.ac.kr

Abstract. Multiple instance learning (MIL) has become a standard paradigm for whole slide image (WSI) analysis in digital pathology, as it enables slide-level prediction without dense annotations. Existing MIL methods typically rely on exhaustive extraction and encoding of high-resolution patches. However, this practice suffers from two critical limitations in real-world clinical settings: it struggles to capture global visual cues at lower magnifications, and incurs substantial computational overhead due to the massive number of high-resolution patches per slide. To address these limitations, we propose an efficient low-resolution multiple instance learning (LRMIL) framework that transfers high-resolution knowledge to low-resolution representations. LRMIL adopts a two-stage distillation strategy. First, patch-level cross-resolution distillation aligns low-resolution patch embeddings with high-resolution representations. Second, slide-level knowledge distillation trains a low-resolution student MIL model under both slide-level supervision and teacher guidance. At inference time, LRMIL operates exclusively on low-resolution patches, substantially reducing data preprocessing and computational cost. Extensive experiments on multiple WSI benchmarks demonstrate that LRMIL consistently outperforms state-of-the-art MIL methods while achieving more efficient inference. These results highlight LRMIL as a practical and scalable solution for WSI analysis in clinical pathology. Code is available at https://github.com/hvcl/LRMIL.git.

Keywords: Whole Slide Image · Multiple Instance Learning · Knowledge Distillation.

1 Introduction

Whole slide image (WSI) analysis is a fundamental task in computational pathology, where learning methods must cope with extremely large images and limited annotation availability. Multiple instance learning (MIL) has therefore become a standard paradigm, as it enables slide-level prediction by aggregating patch-level representations without per-patch annotations. Recent MIL-based

† Corresponding author.

approaches have shown superior results in pathology applications [9,11,14,15,19]. However, most existing methods rely on exhaustive extraction and encoding of high-resolution (HR) patches, which introduce critical limitations in both representation capacity and computational efficiency.

Specifically, reliance on high-magnification patches alone restricts the modeling of global visual cues—such as tissue architecture and spatial context—that are more readily captured at lower magnifications. In addition, a single WSI may contain thousands to tens of thousands of HR patches, and the associated preprocessing overhead, including patch extraction and feature encoding, frequently dominates the overall computational cost of the pipeline. Several recent studies attempt to alleviate this issue by filtering or selecting informative HR patches during MIL inference [4,6,7,16]. However, such strategies ultimately rely on HR patch encoding for final prediction and do not explicitly transfer HR semantic knowledge to low-resolution (LR) representations. As a result, LR patches alone remain insufficient for accurate inference, and the computational benefits of LR processing cannot be fully realized. Consequently, the scalability and practicality of existing HR-based MIL pipelines remain limited in real-world clinical settings.

Recently, knowledge distillation (KD) has emerged as an effective paradigm for improving efficiency while preserving performance, by transferring knowledge from a large teacher to a smaller student model [8]. Originally developed for convolutional neural networks, KD has since been extended to various architectures (*e.g.*, vision transformers) and has shown effectiveness across computer vision tasks [2,3,12,17,18]. In computational pathology, several recent studies have also adopted distillation-based strategies to transfer rich semantic knowledge from large models to more efficient student models [5,20]. However, these approaches mainly focus on pretraining compact or lightweight models through KD, rather than addressing the resolution-dependent inefficiencies of MIL pipelines.

To address these challenges, we propose an efficient low-resolution MIL (LRMIL) framework that decouples training resolution from inference resolution. LRMIL leverages fine-grained information only during training, while enabling inference to be performed exclusively on LR patches. Conceptually, LRMIL consists of two distillation stages: First, we perform patch-level cross-resolution distillation, where a frozen HR patch encoder provides supervision to train an LR encoder with the same architecture. This stage aims to transfer HR semantic knowledge into LR patch representations. Second, an LR-based MIL model is trained using the distilled encoder, guided by both bag-level supervision and slide-level knowledge distillation from an HR-based teacher MIL model. At inference time, only the student MIL model is used, operating solely on LR patches. This design substantially reduces data preprocessing and feature extraction costs for inference WSIs, enabling efficient and scalable deployment in clinical pathology workflows. Our contributions can be summarized as follows:

- We propose an efficient MIL framework for WSIs that is explicitly designed to perform inference using only LR patches.

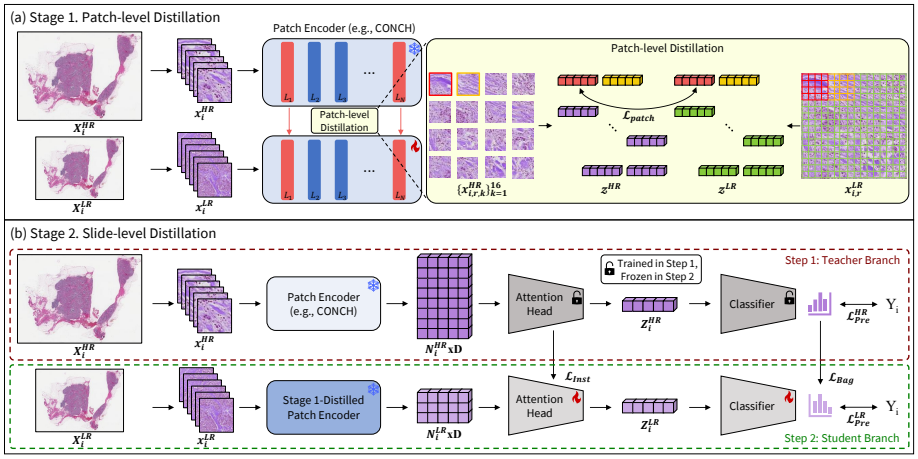

Fig. 1. Overview of our LRMIL framework. (a) Patch-level cross-resolution distillation. Fine-grained semantic knowledge is distilled to a coarse-level patch encoder. (b) Slide-level distillation for MIL. An LR-based student MIL model is trained using both bag-level supervision and teacher guidance.

- We introduce a novel two-stage knowledge distillation strategy, consisting of patch-level cross-resolution distillation and slide-level distillation tailored for computational pathology.
- Through extensive experiments on diverse datasets and tasks, we demonstrate that the proposed method achieves superior performance and computational efficiency compared to state-of-the-art MIL approaches.

2 Methodology

Figure 1 illustrates the overall framework of LRMIL. The key idea of LRMIL is to enable accurate MIL inference using only *low-resolution* patches, while transferring high-resolution knowledge during training through distillation. To this end, LRMIL adopts a two-stage distillation strategy consisting of patch-level cross-resolution distillation (Fig. 1(a)) and slide-level knowledge distillation (Fig. 1(b)). At inference time, only the low-resolution student model is used, enabling efficient WSI analysis without high-resolution processing.

2.1 Stage 1: Patch-level Cross-resolution Distillation

The objective of stage 1 is to transfer fine-grained semantic knowledge from HR observations to LR representations at the patch level, independently of MIL. By aligning matched HR–LR regions, this stage encourages the LR encoder to capture both (i) structural characteristics visible at low magnification and (ii) fine-grained details present at high magnification.

Let X_i^{HR} and X_i^{LR} denote the i-th whole slide image (WSI) at high and low magnifications, respectively. We crop HR and LR patches of the same pixel size (256×256) from X_i^{HR} and X_i^{LR}. In our setting (e.g., $20\times$ for HR and $5\times$ for LR), one LR patch covers the same region as $K = 16$ HR patches. For each LR patch $x_{i,r}^{LR} \subset X_i^{LR}$ covering region r, we therefore identify the corresponding HR patches $\{x_{i,r,k}^{HR}\}_{k=1}^{K} \subset X_i^{HR}$ that jointly cover the same region at higher magnification. This *cross-scale matching* is used to construct distillation pairs.

We adopt a ViT-based architecture for both resolutions. A pre-trained HR encoder $f_{HR}(\cdot)$ is frozen and used as a teacher, while an LR encoder $f_{LR}(\cdot)$ with the same architecture is trained. Our distillation strategy follows prior work in the natural image domain that performs multi-layer distillation between a large teacher and a smaller student model [17]. Specifically, we adopt the same layer selection strategy (e.g., layers $L = \{0, 1, 11\}$ in a 12-layer ViT) and MSE-based alignment objective. However, instead of distilling representations from identical images, we extend this formulation to a cross-resolution setting tailored for computational pathology, where HR and LR patches correspond to the same spatial region at different magnifications.

At each selected layer $l \in \mathcal{L}$, we extract matched HR and LR representations for each spatial region indexed by k. Specifically, for a 256×256 LR patch with a ViT patch size of 16×16 pixels, the encoder produces a 16×16 grid of spatial tokens (256 tokens in total). Each HR patch corresponds to a 4×4 sub-grid (16 tokens) within this LR token grid. Let $\mathcal{S}_k$ denote the index set of LR tokens matched to the k-th HR patch via cross-scale matching. We then define the HR and LR representations at layer l as:

$$z_{i,r,k,l}^{HR} = \mathrm{CLS}_l\big(f_{HR}(x_{i,r,k}^{HR})\big), \quad z_{i,r,k,l}^{LR} = \frac{1}{|\mathcal{S}_k|} \sum_{m \in \mathcal{S}_k} \mathbf{t}_{i,r,m,l}^{LR}, \quad k = 1, \ldots, K, \quad (1)$$

where $\mathbf{t}_{i,r,m,l}^{LR}$ denotes the m-th LR spatial token at layer l, and $|\mathcal{S}_k| = 16$ in our setting. Here, K denotes the number of HR patches matched to a single LR patch, whereas $|\mathcal{S}_k|$ denotes the number of LR spatial tokens pooled for the k-th matched HR patch; although both are 16, they represent different quantities.

We then perform patch-level cross-resolution distillation by aligning each HR CLS embedding with its matched LR region-level token representation. The overall stage 1 objective aggregates losses across the selected layers:

$$\mathcal{L}_{\text{layer}}^{(l)} = \frac{1}{K} \sum_{k=1}^{K} \left\| z_{i,r,k,l}^{HR} - z_{i,r,k,l}^{LR} \right\|_2^2, \quad \mathcal{L}_{patch} = \sum_{l \in L} \mathcal{L}_{\text{layer}}^{(l)}. \quad (2)$$

By distilling HR CLS embeddings into matched LR region-level token representations, stage 1 injects fine-grained HR semantics into spatially corresponding LR features while preserving low-magnification structural cues. The resulting distilled LR encoder is subsequently used to construct LR bags for MIL training in stage 2, without requiring any HR processing at inference time.

2.2 Stage 2: Slide-level Knowledge Distillation

In stage 2, we train an attention-based MIL model that performs accurate bag-level prediction using LR patches only. For this, both the HR patch encoder and the distilled LR encoder from stage 1 are frozen, and only the MIL heads are trained. We optimize the teacher and student MIL models sequentially: train an HR teacher MIL model with bag-level supervision (step 1), and train an LR student MIL model with bag-level supervision and teacher guidance (step 2).

Step 1. Given the i-th WSI X_i^{HR}, we extract and encode HR patches using the frozen HR encoder to obtain instance embeddings. An attention-based MIL head aggregates these embeddings into a bag-level feature $\mathbf{Z}_i^{HR}$ and produces bag-level logits $\mathbf{p}_i^{HR}$. We train the MIL head using only the bag-level label Y_i with the CE loss, $\mathcal{L}_{pre}^{HR} = \mathrm{CE}\big((\mathbf{p}_i^{HR}), Y_i\big)$. Since the encoder is frozen, this step is lightweight and converges quickly in practice.

Step 2. We then train an LR student MIL model on X_i^{LR} using the distilled LR encoder. The student MIL head outputs logits $\mathbf{p}_i^{LR}$, same as step 1. During this step, the teacher MIL model is frozen and provides guidance through both bag-level logits and instance-level attention scores.

Bag-level distillation. We distill the teacher's bag-level prediction by minimizing the KL divergence between the softened distributions:

$$\mathcal{L}_{\text{bag}} = \mathrm{KL}\big(\texttt{softmax}(\mathbf{p}_i^{HR}/T_1) \,\|\, \texttt{softmax}(\mathbf{p}_i^{LR}/T_1)\big)\,, \tag{3}$$

where T_1 is the temperature.

Instance-level distillation. While bag-level distillation aligns slide-level predictions, it does not explicitly enforce patch-level consistency. Therefore, we introduce instance-level distillation to further align attention patterns across resolutions. Let $\boldsymbol{\alpha}_i^{HR}$ and $\boldsymbol{\alpha}_i^{LR}$ denote the attention weights produced by the teacher and student MIL heads, respectively. To align attention across resolutions, we leverage the cross-scale matching defined in stage 1, where each LR patch corresponds to K=16 HR patches covering the same spatial region r. For each LR patch, we construct a region-level teacher attention score, denoted as $\tilde{\boldsymbol{\alpha}}_{i,r}^{HR}$, by averaging the teacher attention weights of the matched HR patches. This averaged score is then used to supervise the corresponding LR patch attention.

(i) Soft attention matching: We match the teacher and student attention distributions using KL divergence on softmaxed scores:

$$\mathcal{L}_{\text{inst}}^{\text{soft}} = \mathrm{KL}\big(\texttt{softmax}(\tilde{\boldsymbol{\alpha}}_i^{HR}/T_2) \,\|\, \texttt{softmax}(\boldsymbol{\alpha}_i^{LR}/T_2)\big)\,, \tag{4}$$

where $\tilde{\boldsymbol{\alpha}}_i^{HR}$ stacks $\tilde{\boldsymbol{\alpha}}_{i,r}^{HR}$ over LR patches, and T_2 is the temperature.

(ii) Hard top-k/bottom-k supervision: To further encourage discrimination, we treat the regions with the highest and lowest teacher scores as pseudo-positive and pseudo-negative instances, inspired by CLAM [14]. Let $\mathcal{P}_i$ and $\mathcal{N}_i$ denote the indices of the top-k and bottom-k elements of $\tilde{\boldsymbol{\alpha}}_i^{HR}$, respectively. We define a binary target $t_{i,r}$=1 for $r \in \mathcal{P}_i$ and $t_{i,r}$=0 for $r \in \mathcal{N}_i$, and optimize:

$$\mathcal{L}_{\text{inst}}^{\text{hard}} = \frac{1}{|\mathcal{P}_i| + |\mathcal{N}_i|} \sum_{r \in \mathcal{P}_i \cup \mathcal{N}_i} \mathrm{BCE}\big(\boldsymbol{\alpha}_{i,r}^{LR},\, t_{i,r}\big)\,, \tag{5}$$

Table 1. Subtype classification results on four datasets. All results except BRCA* correspond to histologic subtype classification. The best result is marked in **bold**.

	Method	TCGA-BRCA			TCGA-NSCLC			TCGA-RCC			BRACS			BRCA*		
		Acc	AUC	t	Acc	AUC	t	Acc	AUC	t	Acc	AUC	t	Acc	AUC	t
Full HR	Max-P	85.2	86.5	49	75.8	84.1	121	91.7	97.5	134	50.6	82.8	27	60.1	78.0	49
	Mean-P	88.5	91.6	49	79.4	89.1	121	92.6	97.8	134	51.1	80.7	27	64.3	80.0	49
	ABMIL	88.3	91.1	49	81.0	89.0	121	92.0	97.5	134	59.5	83.4	27	62.3	81.7	49
	CLAM-MB	86.9	91.6	49	81.9	91.5	121	94.0	98.1	134	**60.3**	83.3	27	63.5	81.9	49
	CLAM-SB	89.6	91.8	49	80.7	89.5	121	93.1	97.0	134	55.3	82.9	27	64.0	82.0	49
	DSMIL	87.5	91.1	49	82.2	90.8	121	93.8	98.7	134	58.8	**83.7**	27	**66.8**	82.6	49
	TransMIL	88.2	92.1	49	81.1	89.6	121	92.9	98.1	135	52.9	83.6	27	61.6	80.5	49
	DTFD-MIL	88.5	91.1	49	83.1	**92.1**	121	93.8	97.9	134	57.3	82.9	27	64.3	82.7	49
SC	ZOOMMIL	88.6	92.2	9	82.4	91.7	22	93.7	98.2	24	57.9	82.5	5	66.2	**82.8**	8
	HDMIL	89.8	91.9	46	83.1	91.0	117	93.1	98.5	133	59.8	82.9	25	65.8	81.8	45
	Ours	**90.7**	**92.3**	**4**	**83.2**	90.6	**10**	**94.2**	**99.0**	**11**	58.8	**83.7**	**3**	**66.8**	**82.8**	**4**

Table 2. Survival prediction results on five datasets. The best result is marked in **bold**.

	Method	TCGA-BRCA			TCGA-LUAD			TCGA-LUSC			TCGA-KIRP			TCGA-KIRC		
		Acc	AUC	t	Acc	AUC	t	Acc	AUC	t	Acc	AUC	t	Acc	AUC	t
Full HR	Max-P	85.1	66.7	49	62.2	60.9	62	56.4	62.5	59	82.8	51.5	45	58.7	71.0	80
	Mean-P	85.7	67.8	49	61.6	60.8	62	57.3	63.1	59	82.9	62.7	45	68.4	70.5	80
	ABMIL	86.2	66.5	49	62.7	61.3	62	58.4	61.6	59	83.1	61.6	45	68.2	69.2	80
	CLAM-MB	85.4	67.3	49	62.9	59.6	62	60.0	63.3	59	83.1	60.5	45	68.3	71.7	80
	CLAM-SB	86.4	67.4	49	61.9	62.0	62	57.5	62.5	59	79.2	61.2	45	68.4	71.2	80
	DSMIL	84.7	65.9	49	61.4	61.3	62	59.2	63.4	59	82.0	59.7	45	68.0	70.4	80
	TransMIL	86.2	64.7	49	57.8	54.7	62	53.9	56.1	59	79.5	55.7	45	66.6	66.3	80
	DTFD-MIL	84.0	64.2	49	62.4	61.3	62	60.6	62.3	59	80.6	61.0	45	68.9	**73.9**	80
SC	ZOOMMIL	86.7	65.3	12	60.5	59.4	12	58.3	62.4	11	82.4	**63.0**	9	68.9	71.3	14
	HDMIL	84.6	66.3	44	62.1	62.0	60	59.9	63.4	55	83.0	62.7	40	68.4	72.7	74
	Ours	**90.0**	**92.0**	**4**	**63.0**	**63.3**	**5**	**61.7**	**63.5**	**5**	**83.3**	60.0	**4**	**70.9**	71.6	**7**

Overall objective. The student MIL model is trained with a weighted sum of bag-level supervision and distillation losses:

$$\mathcal{L}_{\text{slide}} = \mathcal{L}_{pre}^{LR} + \lambda_{\text{bag}}\mathcal{L}_{\text{bag}} + \lambda_{\text{soft}}\mathcal{L}_{\text{inst}}^{\text{soft}} + \lambda_{\text{hard}}\mathcal{L}_{\text{inst}}^{\text{hard}}, \tag{6}$$

where $\mathcal{L}_{pre}^{LR} = \text{CE}((\mathbf{p}_i^{LR}), Y_i)$ and $\lambda_{\text{bag}}, \lambda_{\text{soft}}, \text{and } \lambda_{\text{hard}}$ control contributions of each loss term. At inference time, we discard the HR teacher branch entirely and use only the LR student MIL model, enabling efficient WSI analysis without HR patch extraction or encoding.

3 Experiments and Results

3.1 Experimental Setup

Datasets and Evaluation Metrics. We evaluate LRMIL on three downstream tasks: histologic and molecular subtype classification, and survival prediction.

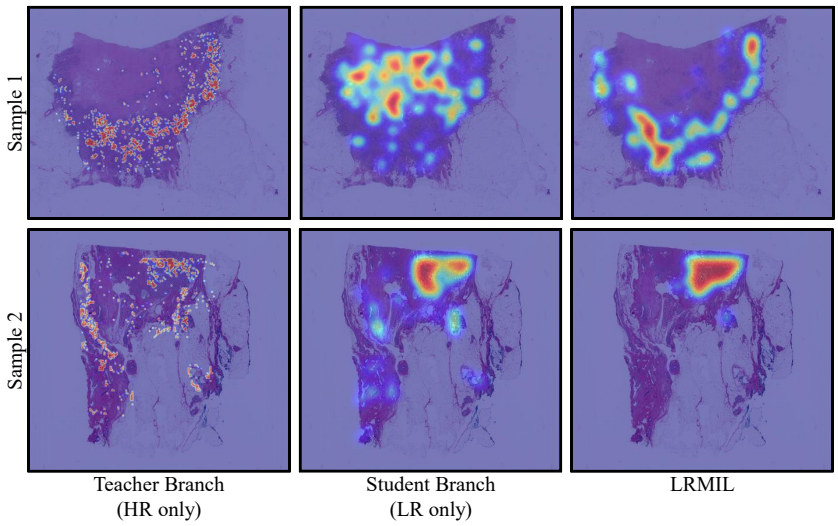

Fig. 2. Visual comparison of attention heatmaps.

For histologic classification, we use four public datasets: TCGA-BRCA (IDC vs. ILC), TCGA-NSCLC (LUAD vs. LUSC), TCGA-RCC (KIRP vs. KIRC vs. KICH), and BRACS (7 classes) [1,10]. For molecular classification, we use TCGA-BRCA to predict LumA, LumB, Basal, and Her2. For survival prediction, we use TCGA cohorts (BRCA, LUAD, LUSC, KIRP, and KIRC) and formulate the task as binary classification (alive vs. dead) based on overall survival status. Note that BRCA* means molecular subtype classification on TCGA-BRCA. We report Accuracy (Acc) and Area Under the ROC Curve (AUC) for all tasks. To evaluate computational efficiency, we measure total inference time t (patch extraction, feature encoding, and MIL aggregation), reported in ($\times 10^4$ s).

Implementation Details. We compare the proposed method with three groups of baselines: (1) conventional pooling methods including max- and mean-pooling (Max-P and Mean-P), (2) state-of-the-art MIL including ABMIL [9], CLAM-MB, CLAM-SB [14], DSMIL [11], TransMIL [15], and DTFD-MIL [19], and (3) efficiency-oriented methods that reduce preprocessing by selectively cropping informative regions at inference time, including ZOOMMIL [16] and HDMIL [4]. For fair comparison, all methods, including HR and LR encoders in LRMIL, use the same CONCH [13] vision encoder as the feature extractor. All models are trained using the Adam optimizer at a learning rate of 1×10^{-4}, early stopping with a patience of 20 epochs, and 5-fold cross-validation. For the slide-level distillation, we set the loss weights to $\lambda_{\text{bag}} = 0.5$, $\lambda_{\text{soft}} = 0.5$, and $\lambda_{\text{hard}} = 0.3$.

3.2 Comparison Results

Quantitative Results. Tables 1 and 2 show the results of subtype classification and survival prediction, respectively. "Full HR" uses all HR patches, while

Table 3. Ablation study of each component in LRMIL. Each dataset corresponds to a different task: histologic subtype classification, molecular subtype classification, and survival prediction. The best result is marked in **bold**.

Loss Function				TCGA-RCC		BRCA*		TCGA-LUSC	
$\mathcal{L}_{patch}$	$\mathcal{L}_{bag}$	$\mathcal{L}_{inst}^{soft}$	$\mathcal{L}_{inst}^{hard}$	Acc	AUC	Acc	AUC	Acc	AUC
				91.4	97.8	62.8	79.8	57.4	62.0
✓				92.6	98.0	63.1	80.7	59.1	62.4
✓	✓			92.8	97.8	65.4	80.9	58.0	**63.7**
✓	✓	✓		**94.2**	98.7	66.2	82.1	61.6	62.8
✓	✓		✓	93.8	98.5	65.6	81.2	61.4	63.0
✓	✓	✓	✓	**94.2**	**99.0**	**67.0**	**82.9**	**61.7**	63.5

"SC" (selective cropping) selects a subset of patches during inference for efficiency. Compared to other methods, LRMIL demonstrates strong performance across all datasets and downstream tasks. Notably, by using exclusively on LR patches, LRMIL reduces total inference time by *more than an order of magnitude* compared to HR-based methods. Furthermore, even compared to efficient "SC" methods, LRMIL achieves both faster inference and superior performance.

Qualitative Results. Figure 2 presents representative attention heatmaps, comparing the HR teacher, LR-only student, and LRMIL. In the first case, the teacher focuses on tumor regions while the LR-only student fails to localize them accurately; in the second case, the LR-only student captures tumor areas that the teacher overlooks. In contrast, LRMIL consistently highlights clinically relevant tumor regions in both cases, suggesting that the model can integrate fine-grained details and global contextual cues across magnifications effectively.

3.3 Ablation Study

Component Ablation. We analyze the contribution of loss terms ($\mathcal{L}_{\text{patch}}$ in stage 1, and $\mathcal{L}_{\text{bag}}$, $\mathcal{L}_{\text{inst}}^{soft}$ and $\mathcal{L}_{\text{inst}}^{hard}$ in stage 2) on representative datasets from each task in Table 3. The comparison between the first and second rows, both without MIL-level distillation, indicates that stage 1 successfully distills essential diagnostic information into a compact representation. Bag-level distillation improves performance over the baseline, and incorporating instance-level supervision further boosts results. Notably, adding $\mathcal{L}_{\text{inst}}^{soft}$ leads to a substantial improvement, underscoring the effectiveness of soft instance-level attention matching. The combination of all three losses achieves the best performance, demonstrating that instance-level alignment provides complementary benefits beyond bag-level matching.

Sensitivity to Top-k. We vary $k \in \{2, 4, 8, 16, 32, 64\}$ and evaluate performance on subtype classification and survival prediction, respectively. The performance remains relatively stable across different values of k, without a clear monotonic trend. This suggests that the proposed hard supervision is robust to the choice of k. In our experiments, we select $k = 4$ as it achieves the best overall performance.

4 Conclusion

In this work, we proposed LRMIL, a low-resolution multiple instance learning framework for efficient whole slide image analysis. By introducing patch-level cross-resolution distillation and slide-level knowledge distillation, LRMIL effectively transfers fine-grained high-resolution semantics to low-resolution representations. Extensive experiments demonstrate that LRMIL achieves superior or competitive performance while substantially reducing inference time. These results highlight the potential of leveraging cross-resolution distillation to enable efficient and scalable WSI analysis in real-world clinical settings. For future work, we plan to extend our framework to more general settings (e.g., regression).

Acknowledgments. This work was supported in part by the National Research Foundation of Korea under Grant RS-2024-00349697 and Grant RS-2021-NR060143; in part by the Institute for Information and Communications Technology Planning and Evaluation under Grant IITP-2026-RS-2020-II201819; in part by the Technology Development Program funded by the Ministry of SMEs and Startups (MSS), South Korea, under Grant RS-2024-00437796; in part by the National Research Council of Science and Technology (NST) grant funded by Korean Government [Ministry of Science and Information and Communications Technology (MSIT)] under Grant GTL24031-000; and in part by Korea University Grant.

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. Chen, G., Choi, W., Yu, X., Han, T., Chandraker, M.: Learning efficient object detection models with knowledge distillation. Advances in neural information processing systems **30** (2017)
3. Chen, P., Liu, S., Zhao, H., Jia, J.: Distilling knowledge via knowledge review. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5008–5017 (2021)
4. Dong, J., Jiang, J., Jiang, K., Li, J., Zhang, Y.: Fast and accurate gigapixel pathological image classification with hierarchical distillation multi-instance learning. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 30818–30828 (2025)
5. Filiot, A., Dop, N., Tchita, O., Riou, A., Dubois, R., Peeters, T., Valter, D., Scalbert, M., Saillard, C., Robin, G., et al.: Distilling foundation models for robust and efficient models in digital pathology. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 162–172. Springer (2025)

6. Guo, H., Zhang, Q., Gao, Z., Yang, S., Peng, S., Tao, X., Yu, T., Wang, Y., Li, Q.: Efficient multi-slide visual-language feature fusion for placental disease classification. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 8018–8027 (2025)
7. Guo, Z., Xiong, C., Ma, J., Sun, Q., Feng, L., Wang, J., Chen, H.: Focus: Knowledge-enhanced adaptive visual compression for few-shot whole slide image classification. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 15590–15600 (2025)
8. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
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. Kefeli, J., Tatonetti, N.: Tcga-reports: A machine-readable pathology report resource for benchmarking text-based AI models. Patterns **5**(3) (2024)
11. 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)
12. Lin, S., Xie, H., Wang, B., Yu, K., Chang, X., Liang, X., Wang, G.: Knowledge distillation via the target-aware transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10915–10924 (2022)
13. 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)
14. 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)
15. 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)
16. Thandiackal, K., Chen, B., Pati, P., Jaume, G., Williamson, D.F., Gabrani, M., Goksel, O.: Differentiable zooming for multiple instance learning on whole-slide images. In: European Conference on Computer Vision. pp. 699–715. Springer (2022)
17. Yang, Z., Li, Z., Zeng, A., Li, Z., Yuan, C., Li, Y.: Vitkd: Feature-based knowledge distillation for vision transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1379–1388 (2024)
18. Yang, Z., Zeng, A., Li, Z., Zhang, T., Yuan, C., Li, Y.: From knowledge distillation to self-knowledge distillation: A unified approach with normalized loss and customized soft labels. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17185–17194 (2023)
19. 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)
20. 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)