SurvMILKD: A Weakly Supervised Survival Analysis Framework for Multi-Teacher Knowledge Distillation using Pathology Foundation Models

Mayur Mallya[1], Ali Khajegili Mirabadi[1], Hossein Farahani[1], and Ali Bashashati[1]

University of British Columbia, Canada
{mayur.mallya,ali.mirabadi,h.farahani,ali.bashashati}@ubc.ca

Abstract. Accurate survival modeling using whole slide images (WSI) is crucial for guiding cancer treatment and improving patient outcomes, yet it remains challenging due to the gigapixel resolution, small data cohorts, and limited annotations. In recent years, pathology foundation models (FM) have shown impressive performance improvements over the previous benchmarks set by the ImageNet-based backbones across a variety of WSI analysis tasks. However, the performance of individual FMs, especially in a secondary analysis task such as survival modeling, varies depending on the dataset, without a clear consensus for a superior FM. With over 25 publicly available pathology FMs so far, in this work, we address the growing challenge of model selection for survival analysis in digital pathology. We propose **SurvMILKD**, the first multi-teacher knowledge distillation (MKD) framework for WSI-based survival analysis. To efficiently enable MKD with high resolution of the WSIs, we integrate the weakly supervised multiple instance learning (MIL) adapters in our framework. The objective of the framework is to distill the complementary task-specific knowledge from multiple teacher FMs into the student FM, which, during inference, has collective knowledge of all FMs involved during training. Additionally, we introduce the risk-aware distillation loss to distill outcome-specific knowledge from multiple teachers into the student model. Our experiments on eight public datasets (including two external evaluation datasets) demonstrate that the proposed approach outperforms the individual FMs, their ensemble, as well as the prior MKD approaches that we benchmarked on survival analysis for the first time in literature. The codes are available at `https://github.com/AIMLab-UBC/SurvMILKD`.

Keywords: Digital Pathology · Knowledge Distillation · Foundation Models

1 Introduction

The histopathology whole slide images (WSI), considered as the gold standard for cancer diagnosis, contain rich, spatially resolved morphological patterns that

reflect the underlying tumor biology, disease aggressiveness, and ultimately patient prognosis. Modeling time-to-event outcomes directly from WSIs is therefore crucial in oncology for enabling precise risk stratification and personalized treatment guidance. However, unlike primary diagnostic tasks such as grading or subtyping, which are guided by established criteria and can be directly observed from tissue morphology, the secondary task of survival analysis necessitates indirectly inferring outcomes from these patterns. Additionally, with the presence of time-censored observations, effective survival analysis requires models capable of learning a continuous risk representation rather than discrete categories, while efficiently aggregating outcome signals across gigapixel-scale WSIs.

In the recent years, foundation models (FM) in pathology have gained prominence, crediting the success to their counterparts in natural language processing [5,12,21] and mainstream computer vision [1,9,14,22]. Pre-trained on millions of WSI patches, FMs such as UNI [7], Prov-GigaPath [59], and Virchow [70] have outperformed the benchmarks previously set by the ImageNet-based backbones across a variety of downstream tasks including tumor subtyping, biomarker prediction, and survival analysis [7, 10, 28, 54]. The widespread success of domain specific FMs led to the surge in pathology FMs with different pre-training cohorts, model architectures, and learning objectives. As of today, there are over 25 distinct pathology FMs in the literature [2].

The early FMs such as CTransPath [53], Phikon [19], and Lunit-DINO [28] leveraged the publicly available WSI cohorts such as The Cancer Genome Atlas (TCGA) [55], with nearly 30,000 WSIs for their pre-training, while, on the other hand, the newer FMs used much larger proprietary datasets with upwards of 100,000 WSIs for pre-training. The use of private data cohorts for pre-training allowed these newer FMs to be evaluated and compared with other FMs on public datasets without any concerns of data leakage. However, despite several benchmarking efforts [4, 40, 42, 56], so far, there has been no clear consensus on a superior or a go-to FM for a given dataset or task. With even more FMs coming out lately [29, 63, 67], the challenge of model selection among pathology FMs is a growing concern.

Furthermore, this shift from publicly available pre-training datasets to large proprietary cohorts introduces additional concerns regarding transparency, such as the patient diversity and quality control standards of the private datasets, making it difficult to assess potential biases in the learned representations. Even among foundation models that disclose their data sources [7, 59, 70], much of the training data is collected from regional hospitals, which may limit geographic and demographic diversity. This homogeneity in the data origins can induce risk of dataset-specific bias, while underscoring the importance of leveraging multiple distinct FMs to obtain robust representations (see the supplementary section for more details).

A traditional yet intuitive strategy to address the model selection problem is to employ an ensemble strategy that combines complementary strengths of multiple FMs to improve overall performance [20, 60, 65]. However, given the large scale of modern pathology FMs (sometimes exceeding 1 billion parameters) this

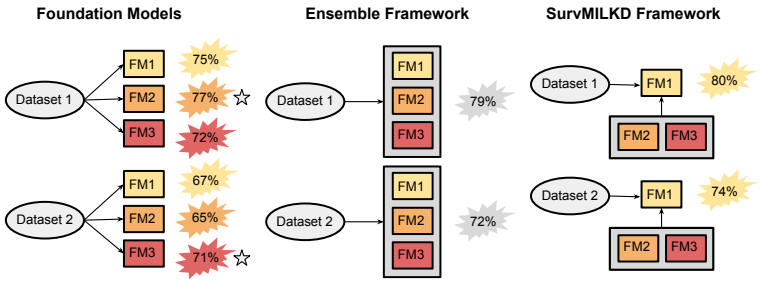

Fig. 1: Foundation models showing different levels of performance across different datasets, leading to the model selection problem. Ensemble framework circumvents this issue, but is bulky at inference. **SurvMILKD** addresses both issues of model selection and bulky inference.

approach quickly becomes computationally impractical. Incorporating additional FMs into an ensemble significantly increases the inference costs, limiting the feasibility of its application for real-world and low-resource settings. Moreover, the high dimensionality of the ensemble representations can introduce optimization difficulties, resulting in an inefficient model.

As an alternative to ensembling, knowledge distillation (KD) offers a more efficient way to combine multiple foundation models (FMs) into a single model [23, 41]. In a multi-teacher knowledge distillation (MKD) framework, knowledge from several teacher FMs is transferred to a single student FM during training. By learning from multiple teachers, the student is encouraged to capture complementary strengths from each model rather than relying on any single representation. Importantly, unlike ensemble methods that require all constituent models at inference time, MKD only requires the trained student model for downstream evaluation. This significantly reduces computational cost while retaining the benefits of multiple teachers, making MKD a scalable and practical solution to the model selection problem.

In this work, we introduce **SurvMILKD**, an MKD framework designed for survival analysis with WSIs, specifically to address the model selection challenge among pathology FMs. Our approach leverages multiple pre-trained histopathology foundation models (FMs) as teachers to guide the task-specific training of a student model. By distilling complementary knowledge from multiple teachers, we aim to learn a unified student representation that captures their collective strengths, with the potential to surpass the individual FMs and the computationally expensive ensemble model. To accommodate the gigapixel scale of WSIs, we adopt the multiple instance learning (MIL) framework under weak supervision, where each slide is decomposed into patches, and the patient-level outcome label, along with supervisory signals from the teachers are propagated to the patch level during training. Our framework advances research at the intersection of MIL, MKD, and WSI survival analysis in several key aspects:

- We propose the first MKD framework for survival analysis with WSIs. By incorporating multiple state-of-the-art pathology FMs as teachers, the proposed **SurvMILKD** student model, along with a trainable MIL adapter, consistently surpasses the performance of individual FMs, their ensemble, along with prior survival and MKD baselines across eight public WSI datasets, including two datasets on external evaluation.
- By adapting the prior MKD techniques originally developed for natural image classification to WSI survival analysis, we establish the first survival MKD benchmark in this domain. To handle the high resolution of WSIs ($50,000 \times 50,000$ average resolution in contrast to 512×512 resolution of natural images), we integrate the MIL paradigm across all the prior works.
- To enable knowledge distillation with scalar risk outputs of the survival models (in contrast to the softmax probability outputs in classification models), we propose the risk-aware distillation loss specifically for knowledge distillation in survival models.
- Through extensive experiments, such as the analyses on the number of teachers, choice of the student model and the MIL adapter, computational cost analysis, and interpretability through attention heatmaps, the proposed framework provides strong empirical evidence of its effectiveness and offers practical guidance for designing task-specific knowledge distillation strategies for MKD in survival analysis with WSIs.

2 Related Works

2.1 Pathology Foundation Models

Over the past three years, more than 25 histopathology FMs have been introduced [2], but despite the recent advancements, the performance on WSI downstream tasks remains highly sensitive to the choice of the backbone feature extractor, highlighting the importance of careful model selection. A majority of these FMs are pre-trained on WSI patches with a vision transformer backbone using different learning objectives, such as the DINO [43] or the masked image modeling [58] frameworks. Among them, H-Optimus-0 [47] and Prov-GigaPath [59] are notable for their scale, each exceeding one billion parameters, while Virchow2 [70] is trained on one of the largest pre-training cohorts, incorporating over three million WSIs. Some FMs integrated the text modality in their pre-training, with models such as CONCH [35], PLIP [24], and QUILT [26] leveraging the image-text pairs for pre-training. More recently, multimodal foundation models have been introduced, incorporating additional supervision from auxiliary modalities such as clinical reports (*e.g.* TITAN [13], PRISM [49]) or molecular data (*e.g.* THREADS [52]). In this work, we use UNI, H-Optimus-0, Prov-GigaPath, Virchow2, and CONCH (v1.5), based on their diverse pre-training sources (more details in the supplementary section) and demonstrated strong performance in external evaluations.

2.2 Survival Analysis with Whole Slide Images

The gigapixel resolution of whole-slide images (WSIs) and the absence of region-level annotations have made multiple instance learning (MIL) a dominant framework for computational pathology. Under weak supervision, a slide is represented as a bag of patches that are aggregated to predict slide-level outcomes. Attention-based MIL methods such as ABMIL [27] and its variants (such as CLAM [36], VarMIL [48], DeepAttnMISL [61], ILRA [57]) have demonstrated strong performance by learning to weight informative regions, while transformer (TransMIL [50], HIPT [6]), graph (PatchGCN [8], MamMIL [18], DAS-MIL [3], WiKG [34]), and multi-magnification (DSMIL [33], ZoomMIL [51]) based extensions further model spatial context within WSIs. In parallel, survival analysis with whole-slide images has evolved from early CNN pipelines that extracted features from sampled regions of interest and fed them into Cox models [68, 69], to more advanced multiple instance and weakly supervised methods that learn to weigh informative tiles via attention or ranking losses, better handling WSI heterogeneity and censoring [32, 66]. In our work, we use ABMIL due to its simplicity and demonstrated robustness across various pathology FMs, making it well-suited for our application that requires multiple MIL adapters on each of FMs involved in MKD.

2.3 Multi-Teacher Knowledge Distillation

Early KD works focused primarily on transferring knowledge from a large teacher model to a smaller student model, but since then, the KD applications have expanded to broader settings, including cross-modal distillation [17, 25, 38] and multi-teacher distillation (MKD) [64]. Similar to conventional KD, MKD methods rely on logit-based or feature-based distillation losses. However, the presence of multiple teachers introduces the additional challenge of aggregating supervisory signals effectively, requiring mechanisms to determine which teachers should contribute more strongly during distillation. Different MKD approaches address this aggregation in various ways, for example, EBKD [31] employs entropy-based teacher weighting, AEKD [15] uses gradient-based criteria, and CAMKD [64] applies confidence-based criteria to weigh the teacher contributions. Methods such as FitNet [46] and CAMKD further incorporate feature similarity losses to align student representations with those of selected teachers. More recently, MKD strategies have also been explored in the context of self-supervised learning for training general-purpose foundation models. For instance, GPFM [37] leverages UNI, CONCH, and Phikon as teachers, while AM-Radio [45] uses CLIP [44], DINOv2 [43], and SAM [30] for additional supervision. However, these approaches rely on extremely large-scale datasets to train general-purpose FMs in a self-supervised manner, unlike our task-specific MKD framework that is trained in a supervised manner, similar to MIL training. On the other hand, HistoMILKD [39] more recently introduced an MKD framework for WSI classification, in contrast to the survival based MKD framework proposed here.

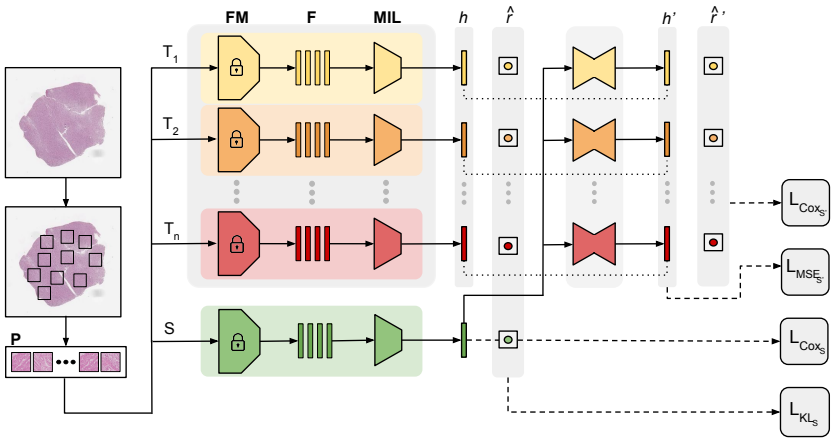

Fig. 2: Overview of the proposed **SurvMILKD** framework during the training phase. The WSI patches (**P**) are passed through all the teacher models (denoted by $T_1, T_2, \ldots, T_n$) and the student model (S) to obtain respective slide-level representations (h) and risk scores ($\hat{r}$). The inputs to the loss functions are guided by the arrows. The lock symbol indicates the FMs are pre-trained and frozen.

3 Methodology

3.1 Problem Formulation

We denote each WSI as a bag of image patches $\mathbf{P} = \{p_1, p_2, \ldots, p_k\}$, where p_j denotes the j-th patch and k is the total number of patches extracted from the WSI. Each patch p_j is processed by a pre-trained FM to obtain a task-agnostic feature embedding. This results in a set of patch-level representations $\mathbf{F} = \{f_1, f_2, \ldots, f_k\}$, where $f_j \in \mathbb{R}^d$ corresponds to the feature vector of patch p_j, and d denotes the embedding dimensionality. In our MKD framework, we denote the student model by subscript S and the teacher models by subscripts $T_1, T_2, \ldots, T_n$, where n is the total number of teachers. Accordingly, $\mathbf{F}_S$ represents the WSI feature embeddings extracted using the student FM, while $\mathbf{F}_{T_1}, \mathbf{F}_{T_2}, \ldots, \mathbf{F}_{T_n}$ denote the corresponding feature embeddings obtained from each teacher FM. These pretrained FM representations serve as fixed inputs to the **SurvMILKD** framework, as all FMs remain frozen during training.

3.2 Model Architecture

We employ five pre-trained FMs: UNI, H-Optimus-0, Prov-GigaPath, Virchow2, and CONCH (v1.5). Among these, UNI has the smallest pre-training dataset and model size; therefore, we designate UNI as the student model and treat the remaining four FMs as teachers in our framework. To aggregate patch-level embeddings into slide-level representations, we adopt ABMIL as the MIL adapter. ABMIL is chosen due to its architectural simplicity, small size, and the strong

empirical performance across diverse downstream tasks and FM backbones. In our setup, we use five independent ABMIL modules (one per FM) to produce slide-level latent representations in their respective embedding spaces, as shown in Fig. 2. Further, to enable knowledge transfer from teachers to the student, we introduce a feature translation module that maps student representations into each teacher's latent space. This module follows an encoder–decoder design implemented using MLP layers. Since we employ four teacher models, we instantiate four independent translation modules, each reconstructing the corresponding teacher representation. During inference, only the student branch is used, ensuring that the final model maintains the computational efficiency of a single foundation model, as shown in Fig. 3.

3.3 Loss Functions

Cox partial log-likelihood loss (L_{Cox_S}): We adopt a weakly-supervised MIL framework to train WSI-based survival models and train an MIL model for each of the teacher FMs. Each MIL model takes the patch-level representation $\mathbf{F}$ of a WSI as input and produces a slide-level latent representation (denoted by h) along with a scalar risk prediction (denoted by $\hat{r}$) for the given WSI. Concretely, for each teacher branch we have

$$\hat{r}_{T_i},\, h_{T_i} = \mathbf{MIL}_{T_i}(\mathbf{F}_{T_i}), \quad \forall i \in \{1, 2, \ldots, n\}, \tag{1}$$

and for the student branch,

$$\hat{r}_S,\, h_S = \mathbf{MIL}_S(\mathbf{F}_S). \tag{2}$$

We train the teacher MIL models $\mathbf{MIL}_{T_1}, \mathbf{MIL}_{T_2}, \ldots, \mathbf{MIL}_{T_n}$ beforehand and use these as initial weights when training the student $\mathbf{MIL}_S$ model. While the teacher models are optimized independently in their respective embedding spaces, the student model is trained using a combination of loss terms, one of which is the Cox loss, denoted by L_{Cox_S}.

Let (t_j, e_j) denote the observed follow-up time and event indicator for patient j, where $e_j \in \{0, 1\}$ indicates whether the event was observed (1) or censored (0). Let $\hat{r}_j$ denote the predicted risk score for patient j (higher indicates worse prognosis), and define the risk set for patient j as $\mathcal{R}(t_j) = \{k \mid t_k \geq t_j\}$. Then, the negative Cox partial log-likelihood for the student is given by

$$L_{Cox_S} = -\frac{1}{\sum_j e_j} \sum_{j:\, e_j=1} \left(\hat{r}_{S,j} - \log \sum_{k \in \mathcal{R}(t_j)} \exp(\hat{r}_{S,k}) \right). \tag{3}$$

Risk-aware distillation loss (L_{KD_S}): Unlike classification, survival models output a continuous scalar risk score rather than a categorical probability distribution. To enable distribution-level distillation via KL divergence, we transform the scalar risk predictions into a soft ordinal distribution over discrete risk levels.

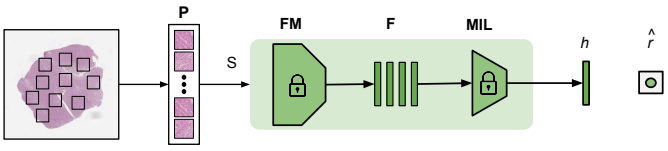

Fig. 3: During inference, **SurvMILKD** only requires the student branch of the framework, without needing any of the bulky teacher branches (as with the ensemble model).

For each teacher T_i, let $\hat{r}_{T_i}$ and $\hat{r}_S$ denote the scalar risk predictions of the teacher and student, respectively. We partition the teacher's risk scores within a batch into four ordinal bins using quartile-based thresholds $\{t_1, t_2, t_3\}$ computed from $\hat{r}_{T_i}$. Instead of hard binning, we use a differentiable relaxation:

$$g_k(r) = \sigma(t_k - r)\,, \quad k \in \{1, 2, 3\}, \tag{4}$$

where $\sigma(\cdot)$ is the sigmoid function (see supplementary section for more details). These cumulative gates are converted into a probability mass over four ordinal risk bins:

$$\mathbf{p}(r) = \big[\, g_1,\ (g_2 - g_1)_+,\ (g_3 - g_2)_+,\ 1 - g_3 \,\big], \tag{5}$$

where $(x)_+ = \max(x, 0)$. This yields a 4-dimensional probability vector $\mathbf{p}(r) \in \mathbb{R}^4$ for each risk score. Applying this transformation to both teacher and student risks produces $\mathbf{p}_{T_i}$ and $\mathbf{p}_S$, respectively. We then perform KL divergence between the softened risk distributions:

$$L_{KD_{(T_i \to S)}} = \mathrm{KL}(\mathbf{p}_{T_i} \,\|\, \mathbf{p}_S)\,, \tag{6}$$

Adaptive teacher weighting. In the multi-teacher setting, not all teachers provide equally reliable survival supervision. We assign weights based on each teacher's batch-wise survival performance. Let $L_{Cox_{T_i}}$ denote the negative Cox partial log-likelihood of teacher T_i computed on the current batch. Teachers with lower Cox loss are considered more reliable. We compute normalized attention weights as:

$$w_{T_i} = \frac{1 - \mathrm{softmax}(L_{Cox_{T_i}})}{n - 1}, \quad \forall i \in \{1, \ldots, n\}. \tag{7}$$

Finally, the overall survival distillation loss for the student is

$$L_{KD_S} = \sum_{i=1}^{n} w_{T_i}\, L_{KD_{(T_i \to S)}}. \tag{8}$$

Feature distillation loss ($L_{MSE_{S'}} + L_{Cox_{S'}}$): To leverage the complementary latent representations learned by different FMs, we encourage the student MIL features (h_S) to reconstruct those of each teacher MIL (h_{T_i}). Specifically, the student feature h_S is passed through a teacher-specific translation module to obtain reconstructed teacher features, denoted by h'_{T_i}. We minimize the discrepancy between the original teacher features and their reconstructions using a mean squared error loss:

$$L_{MSE_{S'}} = \sum_{i=1}^{n} \left\| h_{T_i} - h'_{T_i} \right\|_2^2 . \tag{9}$$

In addition to aligning latent representations, we further supervise the reconstructed teacher features through the survival objective. Let $\hat{r}'_{T_i}$ denote the scalar risk prediction obtained from h'_{T_i}. We apply the Cox partial log-likelihood loss to these reconstructed predictions:

$$L_{Cox_{S'}} = \sum_{i=1}^{n} L_{Cox}(\hat{r}'_{T_i}, t, e), \tag{10}$$

Total loss (L_{total}): During training, we optimize a weighted combination of survival and distillation objectives (from Eqs. 3, 8, 9, and 10). We use hyperparameters α and β to balance the contribution of the distillation losses. The overall training objective is given by

$$L_{total} = L_{Cox_S} + \alpha L_{KD_S} + \beta L_{MSE_{S'}} + L_{Cox_{S'}}. \tag{11}$$

4 Experimental Setup

We evaluate the proposed framework on six public WSI cancer datasets from the TCGA [55] database and two datasets from the CPTAC3 [16] database for external evaluation. Our TCGA datasets include Breast Invasive Carcinoma (BRCA, 1110 slides), Colon Adenocarcinoma (COAD, 421 slides), Head-Neck Squamous Cell Carcinoma (HNSC, 467 slides), Kidney Renal Clear Cell Carcinoma (KIRC, 495 slides), Low Grade Glioma (LGG, 830 slides), and Uterine Corpus Endometrial Carcinoma (UCEC, 548 slides). Our CPTAC3 datasets for external evaluation include HNSC (133 slides) and KIRC (119 slides). For additional details on the datasets, preprocessing steps, training setup, and model optimization settings, please refer to the supplementary section of the paper.

5 Results

5.1 WSI Survival Analysis

We evaluate the performance of **SurvMILKD** framework across eight survival analysis datasets and compare with the individual FMs, their ensemble, as well

Table 1: Results of WSI survival analysis across six TCGA datasets reported using percentage C-index. We compare the performance of **SurvMILKD** with individual FMs, their ensemble, and prior survival and MKD approaches. **Red** and *blue* fonts show the best and second best model, while underline denotes the best performing FM for a given dataset. "Imp." denotes the improvements over the baseline and "p" denotes the p-value evaluated at significance levels of $p < 0.01$ (**) and $p < 0.05$ (*).

Type	Method	TCGA BRCA	TCGA COAD	TCGA HNSC	TCGA KIRC	TCGA LGG	TCGA UCEC	CPTAC3 HNSC	CPTAC3 KIRC
FM	UNI [7]	67.03 ± 1.1	63.02 ± 0.7	57.73 ± 2.3	69.70 ± 0.4	70.61 ± 1.2	64.63 ± 0.6	58.52	55.96
	Virchow2 [70]	65.63 ± 1.6	65.91 ± 1.8	56.49 ± 0.9	70.02 ± 0.8	59.80 ± 1.7	64.30 ± 1.4	56.87	56.77
	Prov-GigaPath [59]	61.60 ± 3.1	65.38 ± 2.3	57.27 ± 1.3	70.16 ± 0.4	69.39 ± 1.9	58.49 ± 0.7	55.71	54.55
	H-Optimus-0 [47]	63.54 ± 3.4	61.90 ± 2.1	56.50 ± 2.0	69.62 ± 1.7	70.91 ± 0.9	66.32 ± 1.4	57.04	55.26
	CONCH (v1.5) [35]	62.83 ± 0.4	59.13 ± 0.8	54.28 ± 1.6	68.97 ± 0.5	68.40 ± 1.6	57.16 ± 3.3	52.63	54.81
	GPFM [37]	67.51 ± 0.8	65.05 ± 1.6	57.95 ± 1.6	70.35 ± 0.7	70.22 ± 1.4	62.77 ± 2.4	57.16	55.48
	Ensemble (5 FMs)	*67.74 ± 0.9*	66.73 ± 1.3	57.86 ± 1.4	**71.11 ± 0.4**	71.35 ± 1.6	65.02 ± 0.5	59.72	58.31
Surv.	DeepAttnMISL [61]	67.53 ± 1.4	67.98 ± 2.0	57.11 ± 2.3	68.78 ± 1.3	70.98 ± 0.8	67.32 ± 1.0	58.88	55.87
	TransMIL [50]	62.11 ± 2.1	61.98 ± 1.6	57.08 ± 1.2	67.68 ± 1.7	70.45 ± 1.3	62.47 ± 1.2	55.34	56.68
	PatchGCN [8]	66.69 ± 1.6	69.54 ± 1.7	57.21 ± 2.2	70.66 ± 1.6	*73.46 ± 1.8*	*66.34 ± 1.5*	57.26	56.81
	WiKG [34]	64.25 ± 1.9	65.01 ± 0.8	55.52 ± 1.4	66.43 ± 0.7	72.23 ± 1.4	63.98 ± 1.0	55.04	55.16
	ILRA [57]	66.52 ± 1.2	70.40 ± 2.2	57.53 ± 1.3	67.75 ± 0.6	69.28 ± 1.1	66.33 ± 1.3	55.44	55.38
MKD	AVER [62]	64.91 ± 1.9	*72.38 ± 1.1*	55.39 ± 0.9	69.93 ± 1.3	71.04 ± 0.3	63.59 ± 0.3	60.98	57.58
	EBKD [31]	63.72 ± 0.7	71.76 ± 0.9	56.73 ± 1.8	69.44 ± 0.3	69.71 ± 2.7	63.36 ± 0.8	60.33	57.58
	AEKD [15]	64.38 ± 1.1	71.67 ± 2.6	56.93 ± 0.4	*70.99 ± 1.1*	73.12 ± 1.1	62.11 ± 1.9	61.23	57.37
	FitNetMKD [46]	66.99 ± 1.5	71.59 ± 0.9	55.60 ± 1.0	70.60 ± 0.7	71.25 ± 1.4	63.32 ± 0.9	59.82	58.79
	CAMKD [64]	65.56 ± 1.5	70.36 ± 0.6	*58.23 ± 0.7*	70.14 ± 0.8	72.69 ± 1.9	64.54 ± 0.8	*61.36*	*59.19*
	SurvMILKD (ours)	**68.32 ± 1.0**	**74.59 ± 0.6**	**58.71 ± 1.5**	70.96 ± 0.2	**74.47 ± 0.6**	**67.38 ± 1.0**	**62.13**	**60.61**
Imp.	over UNI	1.02 ↑	11.57 ↑	0.98 ↑	1.26 ↑	3.86 ↑	2.75 ↑	3.61 ↑	4.65 ↑
	over Ensemble	0.58 ↑	7.86 ↑	0.85 ↑	−0.15 ↓	3.12 ↑	2.36 ↑	2.41 ↑	2.30 ↑
p	Stat. sig. UNI	0.011 *	0.001 **	0.162	0.042 *	0.018 *	0.055	-	-
	Stat. sig. Ensemble	0.279	0.011 *	0.013 *	0.957	0.046 *	0.016 *	-	-

as the prior survival and MKD approaches (originally proposed for natural image classification but were fairly adapted for WSI survival analysis). The individual FMs were benchmarked with the same MIL adapter (ABMIL) and trained using the Cox partial log-likelihood loss as in Eq. (3). Table 1 presents the results of survival analysis across different methods using concordance index (C-index) as the metric, along with their respective model sizes at inference. Each reported value in the table is averaged across 9 experiments (3 folds × 3 seeds) and the values indicate the mean and standard deviation of these experiments. For the external evaluation datasets, we aggregated the risk score of the WSI from different runs using max pooling before computing the C-index.

The first stanza of Tab. 1 shows the performance of different FM backbones used in our framework, when trained individually. Virchow2, H-Optimus-0, and GPFM have the top performance on multiple datasets each while UNI sees top the performance on HNSC external evaluation. CONCH, which was pre-trained with image-text pairs, has a consistently low score in comparison to the other FMs. At the same time, we observe that Virchow2 also has the lowest C-index scores on LGG and BRCA datasets. The lack of clear consensus for the choice of best FM for any given dataset, as seen by the underlined scores dispersed across different FMs, highlights the problem of FM selection for downstream tasks.

Further, we present the ensemble model implemented using the late-fusion strategy [11] which included all five FMs that were benchmarked above. Predictably, the ensemble FM results in a stable performance across all the datasets, outperforming all the individual FMs on five of the six datasets and achieving the second-best score in the case of UCEC dataset. However, the major drawback

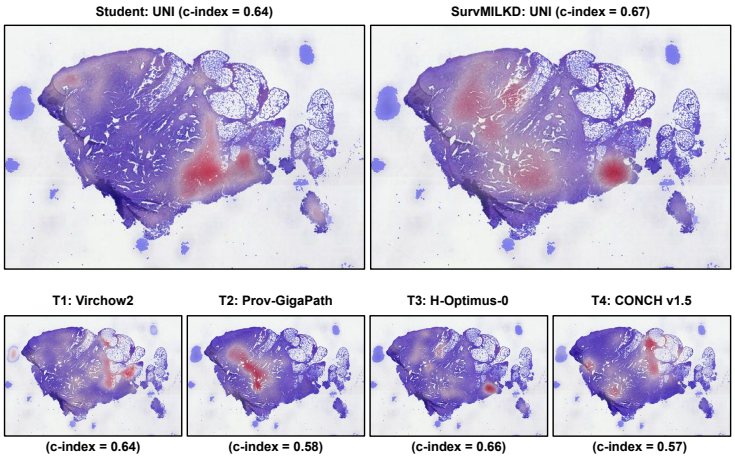

Fig. 4: Attention heatmaps for the UCEC case TCGA-AJ-A23N. Top row compares the heatmaps of the baseline student with the **SurvMILKD** student and the bottom row presents the heatmaps of the teachers (denoted by T1, T2, etc.). For each model, we include the C-index score from Tab. 1 to denote the model performance. We observe that the **SurvMILKD** student attends to regions attended by the baseline student as well as regions attended by some of the teachers (*e.g.* Prov-GigaPath, H-Optimus-0).

of the ensemble model is the model size (3512.6 million parameters), which is more than 10 times the size of UNI and CONCH and three times the size of the largest FMs (Prov-GigaPath and H-Optimus-0). This bulky inference hinders the feasibility of its adoption in real-world settings.

For the prior survival baselines, we use UNI as the backbone feature extraction model before training the models. The MKD approaches, on the other hand, circumvent the issue of model selection and bulky inference by distilling knowledge from the teacher FMs during training. The second stanza in Tab. 1 shows the performance of different MKD approaches. Despite being proposed for natural image classification, these MKD approaches when adapted to WSI survival analysis, achieve better performance than the ensemble FM on four of the six datasets. Finally, the proposed **SurvMILKD** framework achieves the best C-index score on five of the six datasets. In the case of KIRC, **SurvMILKD** is outperformed by the ensemble FM by the smallest margin, as shown in the improvements row (denoted by "Imp."). We observe that in the case of KIRC, the individual FMs have similar performance (ranging from 68.97 to 70.16), possibly benefitting the ensemble over the MKD approach, as the FMs have learned *similar* knowledge, without needing to distill from one FM to another.

Furthermore, we use the one-sided paired t-test to evaluate the statistical significance of our method, which is the **SurvMILKD** UNI model, compared to the baseline UNI model, as well as the ensemble FM model. We observe that

Table 2: Impact of the number of teachers in **SurvMILKD** with UNI as the student backbone. We abbreviate Virchow2, Prov-GigaPath, H-Optimus-0, and CONCH as Vir, PG, H0, and CH, respectively. **Red** and *blue* fonts denote the best and second-best performance per dataset. The numbers represent the percentage C-index scores and ↑ indicates if the student performance improves after knowledge distillation.

Teachers	BRCA	COAD	HNSC	KIRC	LGG	UCEC
UNI (no teachers)	67.03	63.02	57.73	69.70	70.61	64.63
Vir	63.67	67.67 ↑	55.45	68.58	69.11	61.56
PG	61.92	67.19 ↑	58.64 ↑	70.10 ↑	72.01 ↑	57.97
H0	65.63	69.82 ↑	58.07 ↑	70.82 ↑	67.98	62.75
CH	61.10	62.89	53.58	70.67 ↑	72.34 ↑	57.07
Average (1 teacher)	63.08	66.89 ↑	56.44	70.04 ↑	70.36	59.84
Vir + PG	65.11	70.26 ↑	56.88	68.81	71.62 ↑	59.35
Vir + H0	65.70	*74.03* ↑	57.79 ↑	70.77 ↑	71.62 ↑	62.93
Vir + CH	62.00	73.07 ↑	57.41	**71.56** ↑	66.06	64.28
PG + H0	*68.09* ↑	72.81 ↑	57.61	66.98	*73.35* ↑	63.26
PG + CH	66.32	67.80 ↑	56.28	70.48 ↑	71.44 ↑	62.21
H0 + CH	65.84	70.04 ↑	57.75 ↑	70.11 ↑	71.50 ↑	66.12 ↑
Average (2 teachers)	65.51	71.00 ↑	57.29	69.79 ↑	70.93 ↑	63.03
Vir + PG + H0	67.73 ↑	73.25 ↑	57.92 ↑	69.48	73.47 ↑	61.05
Vir + PG + CH	66.16	72.02 ↑	57.87 ↑	70.19 ↑	69.65	*66.27* ↑
Vir + H0 + CH	66.48	70.88 ↑	**58.85** ↑	70.63 ↑	71.08 ↑	63.59
PG + H0 + CH	67.12 ↑	67.99 ↑	57.54	70.33 ↑	73.05 ↑	64.97 ↑
Average (3 teachers)	66.87	71.54 ↑	58.05 ↑	70.16 ↑	71.81 ↑	63.97
Vir + PG + H0 + CH	**68.32** ↑	**74.59** ↑	*58.71* ↑	*70.96* ↑	**74.47** ↑	**67.38** ↑

the proposed framework significantly outperforms the baseline UNI model and the ensemble FM on four datasets each at a significance level of $p < 0.05$.

5.2 Interpretability

In addition to improving the performance on the WSI-based survival analysis, **SurvMILKD**, leveraging the attention mechanism with the ABMIL adapter, allows for the framework to be interpretable. Each patch within a WSI is assigned an attention score that reflects its contribution to the final slide-level risk prediction. Visualizing the high-attention patches can provide insights into the highly influential tissue regions in the model's decision-making process. Figure 4 compares the attention heatmaps of the baseline student model (UNI before MKD) with that of the **SurvMILKD** UNI model (after MKD), and further contrast them with the heatmaps of the teacher FMs used during knowledge distillation. Notably, the attention heatmap after MKD not only retains some of the regions attended by the baseline student model but also includes additional regions attended by the teachers (such as T2: Prov-GigaPath, T3: H-Optimus-0). Interestingly, despite Prov-GigaPath having a lower overall C-index score compared to the other teachers, the strong attention patterns encourage the distilled model to also focus on those regions. This suggests that knowledge distillation not only transfers the risk prediction capability but also the spatial attention patterns from the teachers. For a more detailed analysis including the patient-level risk scores predicted by different models and the failure cases of **SurvMILKD**, please refer the supplementary section.

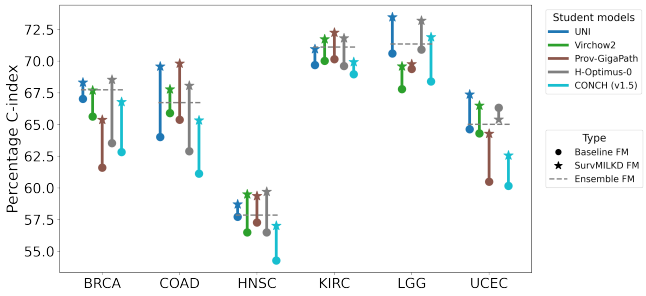

Fig. 5: Analysis of student model choice in **SurvMILKD**. In each case, the student model is guided by the remaining four teacher models.

5.3 Impact of Teacher Models

In Tab. 2, we analyze the impact of the number of teachers on the performance of **SurvMILKD**. In addition to varying the number of teachers, we also evaluate using different combinations of teachers. However, irrespective of the number of teachers used for MKD, we only perform the knowledge distillation once and do not sequentially distill on an already distilled student.

Compared to the baseline UNI model (without teachers), incorporating teachers generally improves survival performance across most datasets. In particular, the 4-teacher configuration achieves the best performance on BRCA, COAD, LGG, and UCEC, while the average 3-teacher and 2-teacher configurations surpasses the baseline UNI on four and three datasets respectively. However, adding more teachers does not always guarantee the best performance. For instance, KIRC achieves the best performance with 2 teachers (Vir + CH), and HNSC with 3 teachers (Vir + H0 + CH).

Interestingly, strong individual teacher performance does not always translate directly to the strongest distilled student. For instance, in the case of COAD, where H0 has a lower performance score than UNI (Tab. 1), we see that using H0 as the teacher leads to the best 1-teacher performance (Tab. 2). Similarly, in several cases, combinations of moderately performing teachers yield superior student results, as in the case of 2-teacher performance in UCEC, with H0 and CH, the FMs with respectively the highest and lowest scores on the dataset (Tab. 1). This behavior suggests that the absolute teacher performance alone does not determine the success of MKD. Instead, the different risk-ranking patterns learned by each FM seem to play a role the distillation process. A deeper analysis of the ranking agreements between the FMs could further clarify these dynamics.

5.4 Impact of Choice of Student Model

While UNI as the student model benefits from **SurvMILKD** distillation, we further analyze whether similar gains can be achieved when other FMs are used

Table 3: Ablation study on the loss functions used in training **SurvMILKD**. **Red** and *blue* fonts show the best and second best percentage C-index scores.

Losses	BRCA	COAD	HNSC	KIRC	LGG	UCEC
L_{Cox_S}	67.03	63.02	57.73	69.70	70.61	64.63
$L_{Cox_S} + \alpha L_{KD_S}$	67.35	71.08	57.64	68.51	*72.28*	63.47
$L_{Cox_S} + \alpha L_{KD_S} + \beta L_{MSE_{S'}}$	*67.56*	*73.25*	*57.95*	*69.84*	71.25	*64.89*
$L_{Cox_S} + \alpha L_{KD_S} + \beta L_{MSE_{S'}} + L_{Cox_{S'}}$	**68.32**	**74.59**	**58.71**	**70.96**	**74.47**	**67.38**

as the student. In conventional knowledge distillation setups, the student model is typically smaller than the teacher models. However, in this analysis, large FMs such as H-Optimus-0 and Prov-GigaPath, each with over one billion parameters, are used as the student, while relatively smaller models such as UNI serve as teachers. Figure 5 presents the results when each FM is alternately used as the student in the SurvMILKD framework. For every configuration, the remaining four FMs act as teachers, and the experiment is repeated in a circular manner (*e.g.* if Virchow2 is the student, the teachers are UNI, H-Optimus-0, Prov-GigaPath, and CONCH). The ensemble FM corresponds to the ensemble of all five FMs.

Across six dataset, **SurvMILKD** consistently improves the performance of the corresponding baseline FM across most datasets, regardless of the choice of student model (with the exception of H-Optimus-0 in the case of UCEC, where the baseline FM had a relatively high performance which dipped after MKD). In many cases, the distilled student also surpasses the ensemble performance (with the general exception of CONCH, where the baseline FM performance is low across multiple datasets that even though MKD improves the performance, it is still lower than that of the ensemble). Overall, this analysis highlights the effectiveness of distilling complementary knowledge from multiple teachers into a single model.

5.5 Loss Ablation

We performed an ablation analysis of the loss components used to train the **SurvMILKD** framework, with results summarized in Tab. 3. As shown in Eq. (11), the overall objective (L_{total}) consists of a weighted combination of four terms: L_{Cox_S}, L_{KD_S}, $L_{MSE_{S'}}$, and $L_{Cox_{S'}}$. In the ablation experiments, we progressively incorporated each additional loss term into the baseline model, which was trained using only the student Cox loss (L_{Cox_S}). Consistent with the experimental setup in Tab. 1, we employed four teacher FMs, used UNI as the student backbone, and adopted ABMIL as the MIL adapter. While the addition of L_{KD_S} alone does not consistently improve the performance across all datasets, the inclusion of $L_{MSE_{S'}}$ on top of L_{KD_S} shows a positive upward trend. However, the full objective yields the best performance indicating that the benefit of survival MKD depends on the appropriate weighting of the loss components. Additional details regarding the influence of the hyperparameters α and β on performance are provided in the supplementary section.

Table 4: Comparison of computational cost across different methods. We abbreviate Virchow2 (Vir2), Prov-GigaPath (PG), H-Optimus-0 (H0), Ensemble (ENS), DeepAttnMISL (DAM), TransMIL (TMIL), PatchGCN (PGCN), FitNetMKD (FMKD). The units M, K and G denote mega, kilo, and giga scales respectively. **Red** and *blue* show the highest and second highest values in the row.

	FM Baselines							Survival Baselines					MKD Baselines					Ours
	UNI	Vir2	PG	H0	Conch	GPFM	ENS	DAM	TMIL	PGCN	WiKG	ILRA	AVER	EBKD	AEKD	FMKD	CAMKD	**SurvMILKD**
Total params (inference) (M)	303.6	631.5	*1135.2*	1135	307.3	303.6	**3512.6**	303.4	304.9	304.9	303.8	303.6	303.6	303.6	303.6	303.9	303.9	303.9
Trainable Params (K)	262.6	655.8	393.7	393.7	197.1	262.6	**1902.9**	139.8	*1454.1*	1383.3	429.4	262.6	262.6	262.6	262.6	682.2	682.2	682.2
GFLOPs (inference)	14.9	37.4	22.4	22.4	11.2	14.9	*167.7*	7.4	**212.6**	126.9	38.3	14.9	14.9	14.9	14.9	14.9	14.9	14.9
GFLOPs (training) / epoch	87.8	218.2	131.1	131.1	65.95	87.8	910.6	64.5	**1050.1**	*1023.2*	746.91	923.6	875.5	875.5	925.7	953.5	953.5	953.5
Peak Memory (GB) w/o FM	0.57	0.68	0.58	0.59	0.55	0.56	**2.85**	0.74	*1.51*	0.69	1.06	0.95	0.92	0.95	1.2	1.1	0.96	0.96

5.6 Computational Cost

In Tab. 4, we compare the computational cost of using **SurvMILKD** (during inference and training) to prior methods such as the FMs, survival and MKD baselines. We observe that **SurvMILKD** requires less resources compared to the ensemble model and some of the prior survival baselines (such as TransMIL and PatchGCN). However, compared to the individual FMs (such as UNI, CONCH, and GPFM), **SurvMILKD** needs more resources during training. During inference, on the other hand, **SurvMILKD** needs the same amount of resources as the student FM (UNI), but leading to better survival analysis performance.

6 Conclusion and Future Work

In this work, we propose **SurvMILKD**, the first MKD framework for survival analysis with WSIs. Given the increasing number of pathology FMs in the literature without clear consensus on a superior FM, our work proposes a timely and efficient solution to aggregate the task-specific knowledge from different FMs, while maintaining the original model size (unlike the ensemble model). We first benchmark the prior MKD works from the natural image classification domain by adapting them to WSI survival analysis by 1) proposing the risk-aware distillation loss to handle knowledge distillation with scalar risk scores and 2) integrating the MIL adapter into the MKD framework to efficiently handle the gigapixel resolution of the WSIs. Our extensive experiments on eight public cancer datasets highlights the utility of the proposed framework by outperforming the individual FMs, their ensemble, and the prior survival and MKD approaches.

Although **SurvMILKD** demonstrates consistent performance gains across multiple datasets, its generalization to large-scale out-of-distribution data requires further investigation. Future works along this domain should evaluate the framework on large, multi-center cohorts to better assess its robustness. Additionally, incorporating complementary modalities such as genomic or textual data through multi-teacher cross-modal distillation may enable richer representations and further expand the utility of the framework.

References

1. Awais, M., Naseer, M., Khan, S., Anwer, R.M., Cholakkal, H., Shah, M., Yang, M.H., Khan, F.S.: Foundation models defining a new era in vision: a survey and outlook. IEEE Transactions on Pattern Analysis and Machine Intelligence **47**(4), 2245–2264 (2025)
2. Bilal, M., Raza, M., Altherwy, Y., Alsuhaibani, A., Abduljabbar, A., Almarshad, F., Golding, P., Rajpoot, N., et al.: Foundation models in computational pathology: A review of challenges, opportunities, and impact. arXiv preprint arXiv:2502.08333 (2025)
3. Bontempo, G., Bolelli, F., Porrello, A., Calderara, S., Ficarra, E.: A graph-based multi-scale approach with knowledge distillation for wsi classification. IEEE Transactions on Medical Imaging **43**(4), 1412–1421 (2023)
4. Breen, J., Allen, K., Zucker, K., Godson, L., Orsi, N.M., Ravikumar, N.: A comprehensive evaluation of histopathology foundation models for ovarian cancer subtype classification. NPJ Precision Oncology **9**(1), 33 (2025)
5. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
6. Chen, R.J., Chen, C., Li, Y., Chen, T.Y., Trister, A.D., Krishnan, R.G., Mahmood, F.: Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16144–16155 (2022)
7. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., Shaban, M., et al.: Towards a general-purpose foundation model for computational pathology. Nature Medicine **30**(3), 850–862 (2024)
8. Chen, R.J., Lu, M.Y., Shaban, M., Chen, C., Chen, T.Y., Williamson, D.F., Mahmood, F.: Whole slide images are 2d point clouds: Context-aware survival prediction using patch-based graph convolutional networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 339–349. Springer (2021)
9. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PmLR (2020)
10. Ciga, O., Xu, T., Martel, A.L.: Self supervised contrastive learning for digital histopathology. Machine learning with applications **7**, 100198 (2022)
11. Cui, C., Yang, H., Wang, Y., Zhao, S., Asad, Z., Coburn, L.A., Wilson, K.T., Landman, B.A., Huo, Y.: Deep multimodal fusion of image and non-image data in disease diagnosis and prognosis: a review. Progress in Biomedical Engineering **5**(2), 022001 (2023)
12. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). pp. 4171–4186 (2019)
13. Ding, T., Wagner, S.J., Song, A.H., Chen, R.J., Lu, M.Y., Zhang, A., Vaidya, A.J., Jaume, G., Shaban, M., Kim, A., et al.: Multimodal whole slide foundation model for pathology. arXiv preprint arXiv:2411.19666 (2024)
14. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is

worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
15. Du, S., You, S., Li, X., Wu, J., Wang, F., Qian, C., Zhang, C.: Agree to disagree: Adaptive ensemble knowledge distillation in gradient space. advances in neural information processing systems **33**, 12345–12355 (2020)
16. Edwards, N.J., Oberti, M., Thangudu, R.R., Cai, S., McGarvey, P.B., Jacob, S., Madhavan, S., Ketchum, K.A.: The cptac data portal: a resource for cancer proteomics research. Journal of proteome research **14**(6), 2707–2713 (2015)
17. Fang, Z., Yang, Y.: Knowledge distillation across vision and language. In: Advancements in Knowledge Distillation: Towards New Horizons of Intelligent Systems, pp. 65–94. Springer (2023)
18. Fang, Z., Wang, Y., Zhang, Y., Wang, Z., Zhang, J., Ji, X., Zhang, Y.: Mammil: Multiple instance learning for whole slide images with state space models. In: 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 3200–3205. IEEE (2024)
19. Filiot, A., Ghermi, R., Olivier, A., Jacob, P., Fidon, L., Camara, A., Mac Kain, A., Saillard, C., Schiratti, J.B.: Scaling self-supervised learning for histopathology with masked image modeling. MedRxiv pp. 2023–07 (2023)
20. Ganaie, M.A., Hu, M., Malik, A.K., Tanveer, M., Suganthan, P.N.: Ensemble deep learning: A review. Engineering Applications of Artificial Intelligence **115**, 105151 (2022)
21. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)
22. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)
23. Hu, C., Li, X., Liu, D., Wu, H., Chen, X., Wang, J., Liu, X.: Teacher-student architecture for knowledge distillation: A survey. arXiv preprint arXiv:2308.04268 (2023)
24. Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T.J., Zou, J.: A visual–language foundation model for pathology image analysis using medical twitter. Nature medicine **29**(9), 2307–2316 (2023)
25. Huo, F., Xu, W., Guo, J., Wang, H., Guo, S.: C2kd: Bridging the modality gap for cross-modal knowledge distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16006–16015 (2024)
26. Ikezogwo, W., Seyfioglu, S., Ghezloo, F., Geva, D., Sheikh Mohammed, F., Anand, P.K., Krishna, R., Shapiro, L.: Quilt-1m: One million image-text pairs for histopathology. Advances in neural information processing systems **36**, 37995–38017 (2023)
27. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning. pp. 2127–2136. PMLR (2018)
28. Kang, M., Song, H., Park, S., Yoo, D., Pereira, S.: Benchmarking self-supervised learning on diverse pathology datasets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3344–3354 (2023)
29. Kather, J., Sainath, S., Lenz, T., Reitsam, N., Foersch, S., Hewitt, K., Wölflein, G., Meneghetti, A., Liang, J., Wolf, F., et al.: Slideflame: A data-efficient vision-language model for anatomically grounded pathology reporting. Research Square (2026)

30. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
31. Kwon, K., Na, H., Lee, H., Kim, N.S.: Adaptive knowledge distillation based on entropy. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 7409–7413. IEEE (2020)
32. Lee, H.H., Santamaria-Pang, A., Merkov, J., Lungren, M., Tarapov, I.: Multi-modal mamba modeling for survival prediction (m4survive): Adapting joint foundation model representations. arXiv preprint arXiv:2503.10057 (2025)
33. 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)
34. Li, J., Chen, Y., Chu, H., Sun, Q., Guan, T., Han, A., He, Y.: Dynamic graph representation with knowledge-aware attention for histopathology whole slide image analysis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11323–11332 (2024)
35. 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)
36. 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)
37. Ma, J., Guo, Z., Zhou, F., Wang, Y., Xu, Y., Li, J., Yan, F., Cai, Y., Zhu, Z., Jin, C., et al.: Towards a generalizable pathology foundation model via unified knowledge distillation. arXiv preprint arXiv:2407.18449 (2024)
38. Mallya, M., Hamarneh, G.: Deep multimodal guidance for medical image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 298–308. Springer (2022)
39. Mallya, M., Mirabadi, A.K., Farahani, H., Bashashati, A.: Histomilkd: A multiple instance learning based multi-teacher knowledge distillation framework for whole slide image classification. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3390–3400 (2026)
40. Mallya, M., Mirabadi, A.K., Farnell, D., Farahani, H., Bashashati, A.: Benchmarking histopathology foundation models for ovarian cancer bevacizumab treatment response prediction from whole slide images. Discover Oncology **16**(1), 1–15 (2025)
41. Mansourian, A.M., Ahmadi, R., Ghafouri, M., Babaei, A.M., Golezani, E.B., Ghamchi, Z.Y., Ramezanian, V., Taherian, A., Dinashi, K., Miri, A., et al.: A comprehensive survey on knowledge distillation. arXiv preprint arXiv:2503.12067 (2025)
42. Neidlinger, P., El Nahhas, O.S., Muti, H.S., Lenz, T., Hoffmeister, M., Brenner, H., van Treeck, M., Langer, R., Dislich, B., Behrens, H.M., et al.: Benchmarking foundation models as feature extractors for weakly-supervised computational pathology. arXiv preprint arXiv:2408.15823 (2024)
43. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
44. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)

45. Ranzinger, M., Heinrich, G., Kautz, J., Molchanov, P.: Am-radio: Agglomerative vision foundation model reduce all domains into one. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12490–12500 (2024)
46. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets. arXiv preprint arXiv:1412.6550 (2014)
47. Saillard, C., Jenatton, R., Llinares-López, F., Mariet, Z., Cahané, D., Durand, E., Vert, J.P.: H-optimus-0 (2024), `https://github.com/bioptimus/releases/tree/main/models/h-optimus/v0`, accessed: 2026-01-15
48. Schirris, Y., Gavves, E., Nederlof, I., Horlings, H.M., Teuwen, J.: Deepsmile: Contrastive self-supervised pre-training benefits msi and hrd classification directly from h&e whole-slide images in colorectal and breast cancer. Medical Image Analysis **79**, 102464 (2022)
49. Shaikovski, G., Casson, A., Severson, K., Zimmermann, E., Wang, Y.K., Kunz, J.D., Retamero, J.A., Oakley, G., Klimstra, D., Kanan, C., et al.: Prism: A multimodal generative foundation model for slide-level histopathology. arXiv preprint arXiv:2405.10254 (2024)
50. 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)
51. 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)
52. Vaidya, A., Zhang, A., Jaume, G., Song, A.H., Ding, T., Wagner, S.J., Lu, M.Y., Doucet, P., Robertson, H., Almagro-Perez, C., et al.: Molecular-driven foundation model for oncologic pathology. arXiv preprint arXiv:2501.16652 (2025)
53. Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Yang, W., Huang, J., Han, X.: Transformer-based unsupervised contrastive learning for histopathological image classification. Medical image analysis **81**, 102559 (2022)
54. Wang, X., Zhao, J., Marostica, E., Yuan, W., Jin, J., Zhang, J., Li, R., Tang, H., Wang, K., Li, Y., et al.: A pathology foundation model for cancer diagnosis and prognosis prediction. Nature **634**(8035), 970–978 (2024)
55. Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J.M.: The cancer genome atlas pan-cancer analysis project. Nature genetics **45**(10), 1113–1120 (2013)
56. Wölflein, G., Ferber, D., Meneghetti, A., Nahhas, O., Truhn, D., Carrero, Z., Harrison, D., Arandjelović, O., Kather, J.: Benchmarking pathology feature extractors for whole slide image classification; 2024 (2024)
57. 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)
58. Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H.: Simmim: A simple framework for masked image modeling. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9653–9663 (2022)
59. Xu, H., Usuyama, N., Bagga, J., Zhang, S., Rao, R., Naumann, T., Wong, C., Gero, Z., González, J., Gu, Y., et al.: A whole-slide foundation model for digital pathology from real-world data. Nature pp. 1–8 (2024)
60. Yang, Y., Lv, H., Chen, N.: A survey on ensemble learning under the era of deep learning. Artificial Intelligence Review **56**(6), 5545–5589 (2023)

61. Yao, J., Zhu, X., Jonnagaddala, J., Hawkins, N., Huang, J.: Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. Medical image analysis **65**, 101789 (2020)
62. You, S., Xu, C., Xu, C., Tao, D.: Learning from multiple teacher networks. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1285–1294 (2017)
63. Zhang, D., Gong, Z., Pang, X., Liu, J., Lu, J., Cui, H., Ge, J., Zeng, Z., Yi, K., Li, Y., et al.: Care: A molecular-guided foundation model with adaptive region modeling for whole slide image analysis. arXiv preprint arXiv:2602.21637 (2026)
64. Zhang, H., Chen, D., Wang, C.: Confidence-aware multi-teacher knowledge distillation. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4498–4502. IEEE (2022)
65. Zhao, M., Ye, N.: High-dimensional ensemble learning classification: An ensemble learning classification algorithm based on high-dimensional feature space reconstruction. Applied Sciences **14**(5), 1956 (2024)
66. Zhou, J., Tang, J., Zuo, Y., Wan, P., Zhang, D., Shao, W.: Robust multimodal survival prediction with conditional latent differentiation variational autoencoder. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10384–10393 (2025)
67. Zhou, X., Sun, L., He, D., Guan, W., Wang, G., Wang, R., Wang, L., Yuan, X., Sun, X., Zhang, Y., et al.: Knowledge-enhanced pretraining for vision-language pathology foundation model on cancer diagnosis. Cancer Cell (2026)
68. Zhu, X., Yao, J., Huang, J.: Deep convolutional neural network for survival analysis with pathological images. In: 2016 IEEE international conference on bioinformatics and biomedicine (BIBM). pp. 544–547. IEEE (2016)
69. Zhu, X., Yao, J., Zhu, F., Huang, J.: Wsisa: Making survival prediction from whole slide histopathological images. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7234–7242 (2017)
70. 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)