8. Appendix

8.1. Ablation Study

We verify the performance of RankMix through ablation study. In addition to observing from Tab. 2 that self-training is helpful for training the model, it will be interesting to examine how the score function learned under different pre-trained models, as described in Sec. 1.2 and Sec. 4.6, affects RankMix. Here, the pre-trained models with diverse performances used as the teacher models can be obtained from the dataset, WSI-Usability, with serious class imbalance. The results are shown in Tab. 4. Our experimental results indicate that RankMix with two-stage training obviously can boost the performance (averaged Accuracy and AUC) of student models, and the improvement is proportional to the performance of the corresponding teacher models.

Besides, in Sec. 4.3, the number k of features used in RankMix can be an arbitrary positive integer and we examine the impact of k on RankMix in Tab. 5, which indicates that $k = min(m(a), m(b))$ exhibits the best overall performance.

On the other hand, recalled from Sec. 4.6 that we want to transfer knowledge from the general MIL methods to boost RankMix. Since there exist several knowledge transfer methods, including self-training [40,45], knowledge distillation [23, 43], and BERT-based models [13, 27, 29], we examine in Tab. 6 how they affect the student model performance. We can find that, among the knowledge transfer methods used for comparison, RankMix consistently obtains better results in terms of averages Accuracy and AUC.

The details are specifically described in the following. First, we introduce three kinds of distillations applied on different outputs.

1-1) knowledge distillation with soft labels [3, 8, 40, 45]: Use a fixed teacher model to generate soft labels to guide the student model as:

$$\mathcal{L}_{distill} = -(\hat{Y}_s/\tau)\log(\hat{Y}_t/\tau), \tag{20}$$

where $\hat{Y}_t$ and $\hat{Y}_s$ are the bag predictions (Eq. (16)) of the teacher model and student model, respectively, τ denotes the temperature parameter, and BCE, instead of cross-entropy, is used to deal with the multi-class problem. We denote this kind of distillations as "BagBCE".

1-2) knowledge distillation with output distributions [23, 43]: Use a fixed teacher model to guide the student model by mimic the output distribution as:

$$\mathcal{L}_{distill} = KL(\hat{Y}_t/\tau, \hat{Y}_s/\tau), \tag{21}$$

where KL denotes the Kullback-Leibler divergence. We denote this kind of distillations as "BagKL".

1-3) knowledge distillation with score functions: Use a fixed teacher model to guide the student model by mimic the score function as:

$$\mathcal{L}_{distill} = KL(\hat{y}_t/\tau, \hat{y}_s/\tau), \tag{22}$$

where $\hat{y}_t := \{\hat{y}_{t,j}|j = 1, \dots m\}$ and $\hat{y}_s := \{\hat{y}_{s,j}|j = 1, \dots m\}$. We denote this kind of distillations as "ScoreKL".

Second, we describe how the distillation losses described above are combined with the original model loss $\mathcal{L}$ (Eq. (14)).

2-1) General knowledge distillation [3, 8, 23, 43, 45]: The entire loss function L_s can be defined as:

$$\mathcal{L}_s = (1-\alpha)\mathcal{L} + \alpha\mathcal{L}_{distill}. \tag{23}$$

2-2) Additional distillation head (DH): In a recent study [40], an additional distillation head will be added to the model for different tasks to avoid the interference of origin loss $\mathcal{L}$ with the distillation loss $\mathcal{L}_{distill}$. The loss function is the same as expressed in Eq. (23) because the number of losses is the number of tasks.

So far, as we can see from Tab. 6, the student models, except for the ones denoted as +RankMix, are obtained from the first two steps.

Finally, we introduce a mechanism of fine-tuning the teacher model without modifying the loss function or requesting distillation head.

3) Fine-tuning methods: In addition to presenting mixup of ranked features in the pre-trained model during the first stage of training, following fine-tuning in [13, 27, 29], RankMix employs knowledge transfer to further fine-tune the pre-trained model. It is neither required to modify loss functions nor required distillation head. This make our method be easily plugged into existing MIL approaches. As shown in Tab. 6, although RankMix does not guarantee to improve performance significantly for all cases, it indeed carries improvements in particular for the teacher models (like FRMIL) that perform good.

Methods	Teacher model		Student model	
	Accuracy	AUC	Accuracy	AUC
DSMIL [28] + RankMix	-	-	90.27%	87.07%
	88.50%	70.56%	92.04%	74.30%
	86.73%	73.83%	86.73%	82.71%
	76.11%	86.60%	90.27%	88.16%
FRMIL [10] + RankMix	-	-	62.83%	93.11%
	92.92%	58.10%	76.11%	75.70%
	83.19%	72.74%	83.19%	76.01%
	83.19%	87.69%	93.81%	93.61%

Table 4. Impact of different teacher models obtained from WSI-Usability with class imbalance on RankMix. The symbol '-' means that we train the model without self-training.

Method/Dataset	Camelyon16			WSI-usability			TCGA-Lung		
	ACC	AUC	AUPRC	ACC	AUC	AUPRC	ACC	AUC	AUPRC
DSMIL + RankMix (k = 2)	89.15	92.07	91.70	90.27	88.16	**28.51**	94.29	97.77	97.48
DSMIL + RankMix (k = 100)	89.15	92.40	91.98	87.61	**88.47**	27.84	94.29	97.78	97.49
DSMIL + RankMix (k = 1000)	89.15	92.37	91.96	89.38	87.23	28.01	94.29	97.79	97.50
DSMIL + RankMix (default)	**89.92**	**93.47**	**92.74**	**90.27**	88.16	28.41	**94.29**	**98.04**	**97.79**
FRMIL + RankMix (k = 2)	90.70	94.57	93.85	80.53	85.20	43.17	90.48	95.40	95.47
FRMIL + RankMix (k = 100)	89.15	**94.62**	93.71	71.68	76.32	31.48	89.05	95.04	95.11
FRMIL + RankMix (k = 1000)	89.92	94.57	93.67	88.50	67.76	25.65	90.48	95.51	94.96
FRMIL + RankMix (default)	**91.47**	94.59	**93.99**	**93.81**	**93.61**	**43.65**	**93.33**	**97.00**	**97.04**

Table 5. RankMix with different feature number k.

Models/Datasets		Camelyon16		WSI-usability		TCGA-Lung	
		Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
Teacher model	DSMIL [28]	86.82%	93.32%	76.11%	86.60%	93.81%	97.89%
Student models	DSMIL + BagBCE	88.37%	92.22%	87.61%	86.45%	94.29%	98.06%
	DSMIL* + BagBCE	89.92%	91.96%	90.27%	88.32%	94.29%	97.76%
	DSMIL + BagBCE + DH	88.37%	91.76%	90.27%	86.92%	94.29%	**98.08%**
	DSMIL* + BagBCE + DH	89.15%	91.94%	89.38%	87.07%	94.29%	97.70%
	DSMIL + BagKL	88.37%	92.07%	87.61%	87.38%	94.29%	97.99%
	DSMIL* + BagKL	89.15%	91.86%	89.38%	86.92%	94.29%	97.80%
	DSMIL + BagKL + DH	84.50%	92.14%	90.27%	86.14%	94.29%	97.98%
	DSMIL* + BagKL + DH	89.15%	92.65%	90.27%	88.16%	93.81%	97.86%
	DSMIL + ScoreKL	83.72%	91.66%	89.38%	88.16%	94.29%	98.05%
	DSMIL* + ScoreKL	89.15%	92.19%	90.27%	**88.79%**	94.29%	97.77%
	DSMIL + ScoreKL + DH	83.72%	91.73%	90.27%	86.14%	94.29%	98.00%
	DSMIL* + ScoreKL + DH	89.15%	92.24%	90.27%	88.47%	94.29%	97.76%
	DSMIL + RankMix	**89.92%**	**93.47%**	**90.27%**	88.16%	**94.29%**	98.04%
Teacher model	FRMIL [10]	89.15%	94.57%	83.19%	87.69%	90.95%	95.38%
Student models	FRMIL + BagBCE	87.60%	92.83%	69.03%	84.74%	93.33%	96.96%
	FRMIL* + BagBCE	89.92%	94.52%	91.15%	68.69%	89.05%	95.15%
	FRMIL + BagBCE + DH	89.15%	92.78%	73.45%	81.15%	92.86%	96.38%
	FRMIL* + BagBCE + DH	91.47%	94.54%	59.29%	81.46%	90.00%	94.90%
	FRMIL + BagKL	89.92%	94.36%	74.34%	84.42%	93.33%	96.82%
	FRMIL* + BagKL	89.92%	94.46%	90.27%	76.01%	90.00%	94.58%
	FRMIL + BagKL + DH	89.92%	94.26%	72.57%	85.36%	91.90%	95.92%
	FRMIL* + BagKL + DH	90.70%	94.18%	84.96%	73.21%	89.52%	94.58%
	FRMIL + ScoreKL	86.05%	92.27%	67.27%	66.36%	93.33%	**97.22%**
	FRMIL* + ScoreKL	90.70%	94.36%	92.92%	69.47%	90.00%	95.25%
	FRMIL + ScoreKL + DH	87.60%	92.60%	86.73%	69.31%	92.86%	96.67%
	FRMIL* + ScoreKL + DH	90.70%	94.46%	76.99%	85.05%	89.05%	95.05%
	FRMIL + RankMix	**91.47%**	**94.59%**	**93.81%**	**93.61%**	**93.33%**	97.00%

Table 6. Results on different knowledge transfer techniques with different parameter settings. "*" denotes the student model was initialized with the weights of the teacher model. BagBCE, BagKL, and ScoreKL denote three kinds of knowledge distillations, described in Eq. (20)∼Eq. (22), respectively. "DH" denotes an additional distillation head that is plugged into the student model for predicting the distillation output.