Multi-task Screening for Cervical Diseases via Feature Routing and Asymmetric Distillation

Haotian Jiang[1†], Haolin Huang[1†], Jiangdong Cai[1], Mengjie Xu[1], Zhenrong Shen[2], Manman Fei[2], Xinyu Wang[1], Lichi Zhang[2(✉)], and Qian Wang[1,3(✉)]

[1] School of Biomedical Engineering & State Key Laboratory of Advanced Medical Materials and Devices, ShanghaiTech University, Shanghai, China
qianwang@shanghaitech.edu.cn
[2] School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China
lichizhang@sjtu.edu.cn
[3] Shanghai Clinical Research and Trial Center, Shanghai, China

Abstract. Cervical diseases present a significant global health challenge, especially in resource-limited regions with scarce specialized healthcare. Traditional analysis methods for thin-prep cytologic tests and whole slide images are hindered by their reliance on time-consuming processes and expert knowledge. Although AI-driven approaches have advanced single-task screening, they often face difficulties adapting to multi-task workflows and handling extreme class imbalance, thereby limiting their practical deployment in real clinical settings. To address these challenges, we propose a novel framework, MECDS, for multi-task early screening of cervical diseases. Specifically, we design dynamic feature routing to prevent inter-task interference and selectively process task-relevant features. Furthermore, we employ asymmetric attention levels during knowledge distillation to address class imbalance, thus enhancing performance across diverse classes. Our extensive experiments on a large-scale dataset comprising 29,774 whole slide images demonstrate that MECDS surpasses existing single-task and multi-task models across three key screening tasks: cervical cancer, candidiasis, and clue cell detection. Additionally, MECDS exhibits remarkable extensibility, allowing for the efficient integration of novel diagnostic tasks without the need for exhaustive retraining. This unified framework holds great promise for improving comprehensive screening programs in resource-constrained healthcare environments, potentially advancing early detection and improving health outcomes. Our code is released at Github.

Keywords: Cervical Disease Screening · Multi-Task Learning · Feature Routing · Knowledge Distillation

1 Introduction

Cervical diseases are widely recognized as prevalent and serious issues in gynecological health, affecting millions of women worldwide, particularly in resource-limited regions [1–3]. Fortunately, the progression of cervical cellular alterations

† These authors contributed equally to this work.

can be detected through early screening methods like cytological examination, enabling timely therapeutic interventions [4, 5]. The thin-prep cytologic test (TCT) has become the preferred technique for cervical cytology (CC) screening, which is a widely implemented diagnostic modality [6]. TCT not only facilitates the identification of cervical cancer lesions but also enables the detection of various pathogenic infections that significantly impact patients' reproductive health and quality of life, such as candida colonization and bacterial vaginosis (characterized by the presence of clue cells) [7,8]. However, the current paradigm of CC screening relies on manual microscopic evaluation by pathologists, while the large-scale nature of Whole Slide Images (WSIs) presents significant challenges in terms of time-intensive analysis and demands sophisticated professional expertise [9]. These constraints impede the widespread implementation of CC screening initiatives.

Recent advances in artificial intelligence (AI)-driven computer-assisted diagnosis (CAD) have demonstrated remarkable efficacy across CC screening, especially for individual tasks. Numerous studies have proposed task-specific architectures to improve patch-level analysis for cervical cancer [10] and candidiasis screening [11]. Furthermore, some research has optimized the multi-stage [12] or Multiple Instance Learning (MIL) frameworks [13] to better align with the unique characteristics of TCT WSIs [14]. However, these task-specific models face significant challenges in clinical practice. First, the current single-task models can only diagnose through sequential diagnostic workflows for the multiple items in Cervical cancer screening, which presents critical limitations in efficiency and clinical deployment. Secondly, the inherent class imbalance problem poses a fundamental learning barrier, substantially impeding models' ability to learn discriminative feature representations. Therefore, there exists an imperative need for the development of a unified architectural framework capable of multi-task CC screening.

To tackle the aforementioned issues, we propose a novel **M**ulti-task **E**arly **C**ervical **D**isease **S**creening (**MECDS**) framework, as shown in Fig. 1. The key innovation of MECDS is the incorporation of an **Multi-task Feature Adaptation** strategy (Fig. 1(a)). This strategy not only maintains model extensibility by constraining inter-task interactions while preserving performance metrics, but also dynamically selects specific features from redundant WSI representations for each task, thereby enhancing both model performance and computational efficiency. Additionally, MECDS employs an **Asymmetric Knowledge Distillation** training scheme (Fig. 1(d)) to address the inherent class imbalance characteristic of early screening scenarios. Our experiments demonstrate that MECDS outperforms existing single-task and multi-task methods in three common cervical disease screening tasks: cervical cancer screening, candidiasis detection, and clue cell diagnosis. Moreover, MECDS exhibits remarkable extensibility to novel diagnostic tasks through efficient fine-tuning without the need for retraining.

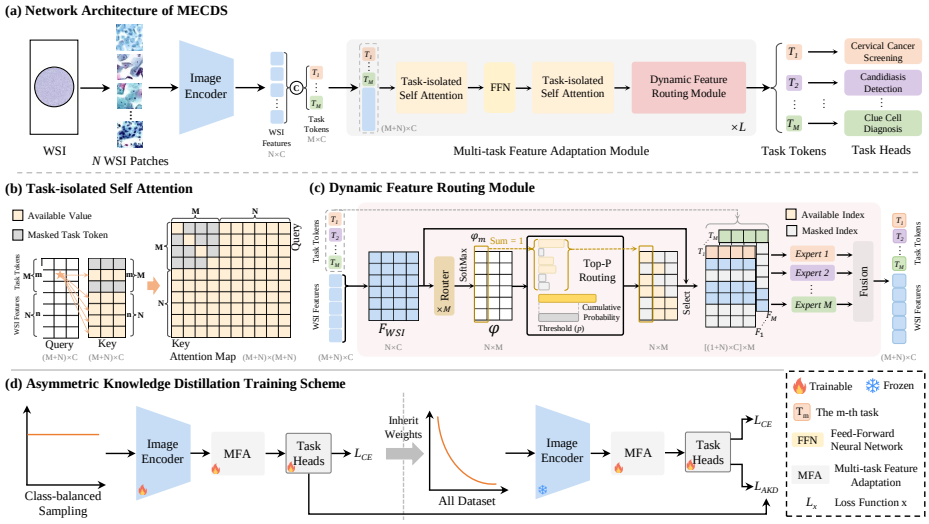

Fig. 1. Overview of the proposed MECDS. (a) The overall architecture of MECDS consists of three main components: the image encoder, multi-task blocks, and task-specific heads. The multi-task learning block main includes (b) Task-isolated Self Attention layers and (c) Dynamic Feature Routing modules. (d) The Asymmetric Knowledge Distillation Training Scheme is introduced to address the extreme class imbalance.

2 Method

2.1 Network Architecture

In the proposed MECDS framework (Fig. 1(a)), each WSI is divided into N patches. These patches are encoded using a LeViT-based [15] image encoder, which is a robust and lightweight architecture that enhances inference speed while maintaining high accuracy. The resulting features from each patch are concatenated to form the WSI's feature $F_{\mathrm{WSI}} \in \mathbb{R}^{N\times C}$, where C represents the dimensions of each patch feature. For multi-task learning, the framework initializes M task-specific tokens $T_m \in \mathbb{R}^{1\times C} (m = 1, 2, ..., M)$ for prediction. These tokens are combined with F_{WSI} and processed through L novel Multi-task Feature Adaptation (MFA) modules. This core component comprises two Task-isolated Self Attention (TSA) layers, a feed-forward network (FFN), and a Dynamic Feature Routing (DFR) module. The TSA layer ensures independent learning processes for each task, preventing inter-task interference and enabling efficient scaling to new tasks. The DFR module dynamically selects task-relevant features from the redundant WSI feature representations based on task-specific requirements. Finally, the updated task tokens are individually processed by corresponding task heads to generate classification results.

2.2 Multi-task Feature Adaptation

Task-isolated Self Attention. For a general early screening model, the ability to efficiently extend to new tasks is essential as it can reduce model maintenance costs. However, traditional models typically demonstrate limited extensibility due to the substantial training costs incurred by re-training the entire model. To address above challenges, we introduce the TSA mechanism shown in Fig. 1(b), which maintains task independence without compromising performance to reduce inter-task interference. Specifically, The TSA mechanism modifies the conventional Transformer's full self-attention architecture [16], where unrestricted token interactions during attention map computation can lead to inter-task interference. Our approach constrains the m-th task token, T_m, to attend only to itself and patch features during attention map calculation, preventing interactions with other task tokens. Thus, when new tasks are introduced, since the new task tokens do not interact with existing task tokens during computation, they do not affect the performance of existing tasks, thereby achieving efficient model extensibility.

Dynamic Feature Routing. WSIs contain rich but often redundant information, making comprehensive processing computationally inefficient. In clinical practice, pathologists optimize their diagnostic workflow by focusing on task-relevant regions, such as specific biomarkers associated with particular diagnoses. Inspired by this clinical selective attention approach, we propose DFR module, illustrated in Fig. 1(c), which adaptively selects task-relevant features from the WSI representation.

The DFR module employs an importance-driven mechanism to dynamically route features based on their relevance to each task. For WSI feature F_{WSI}, we compute a task-relevant matrix $\varphi \in \mathbb{R}^{N \times M}$: $\varphi = \text{Softmax}(\text{Routers}(F_{\text{WSI}}))$, where Routers comprises M learnable MLPs and $\varphi(n, m)$ represents the importance score of the n-th patch feature for the m-th task. According to φ, we implement an adaptive feature selection strategy to select the most relevant features for each task. Specifically, for the m-th task, we select the minimal number of top-ranked features from the descending-sorted φ_m such that their cumulative probability scores exceed the threshold $p = 0.5$, and mask other redundant features. The filtered patch features F_m are considered to be closely associated with the m-th task and adequate for accurate diagnosis. They are sequentially concatenated with the corresponding task token T_m and processed by the m-th expert network Exp_m: $[T_m^{'}, F_m^{'}] = Exp_m([T_m, F_m])$, with $[\cdot]$ denoting the concatenation operation. Notably, in the DRF module, each expert is dedicated to handling a specific task. All processed task-relevant features $F_m^{'}$ are element-wise added to obtain the final feature representation $F_{\text{WSI}}^{'}$, which is then concatenated with all task tokens for subsequent learning.

2.3 Asymmetric Knowledge Distillation

In cervical early screening, datasets typically exhibit extreme class imbalance, with positive samples substantially underrepresented compared to negative ones. This uneven distribution adversely affects both training convergence and test set generalization. To address these issues, we propose an Asymmetric Knowledge Distillation (AKD) training scheme based on the effective distillation [17] framework (Fig. 1(d)). AKD aims to learn generalizable representations while facilitating positive sample learning through a teacher-student paradigm [17,18].

Our training method comprises two distinct phases. In the initial teacher phase, we train the model on a class-balanced subset created through dataset resampling with a cross-entropy loss L_{CE}. This re-sampling training approach may lead the teacher model to overemphasize positive class, potentially compromising overall representation learning [19, 20]. Therefore, in the subsequent student phase, the model is expected to learn generalizable representations by training on a larger imbalanced dataset while distilling positive knowledge from the teacher model. During the distillation process, we adopt an asymmetric strategy that transfers knowledge from the teacher model with more focus on positive samples, thereby mitigating the adverse effects of imbalanced class distribution. Inspired by focal loss [21], We assign different focusing levels, denoted as ω_+ for positive samples and ω_- for negative samples, through asymmetric weights respectively:

$$\begin{cases} \omega_+ = (1-p)^{\gamma_+}, \\ \omega_- = (\hat{p_\alpha})^{\gamma_-}, \end{cases} \tag{1}$$

where $\gamma_+ = 1$ and $\gamma_- = 4$ are focusing parameters for positive and negative samples, respectively. Here, p represents the positive class probability predicted by the student model, and $\hat{p_\alpha}$ is the shifted probability defined as $\hat{p_\alpha} = \max(\hat{p} - \alpha, 0)$, with $\hat{p}$ being the positive class probability predicted by the teacher model and $\alpha = 0.2$ being the probability margin. This asymmetric weight reduces negative samples' contribution when their probability is low. The asymmetric distillation loss L_{AKD} is formulated as:

$$L_{\mathrm{AKD}} = T^2(\omega_+ \hat{z} \log \frac{\hat{z}}{z} + \omega_- \hat{z} \log \frac{1-\hat{z}}{1-z}), \tag{2}$$

where $T = 2$ is the temperature parameter, and $\hat{z} = softmax(\frac{\hat{p}}{T})$ and z are the soft probabilities for positive class predicted by the teacher and student models, respectively. Finally, the student stage is optimized using a composite loss function that combines L_{AKD} and the cross-entropy loss L_{CE}.

3 Experiments

3.1 Dataset and Experimental Setup

Dataset. We establish a large-scale cervical multi-task screening dataset, consisting of 29,774 WSIs with dimensions of $20,000 \times 20,000$ pixels. These images

are collected from multiple scanning devices. Each WSI is annotated for three tasks: cervical cancer screening, candidiasis detection, and clue cell diagnosis. Due to the focus on early screening, each task uses binary class labels (0: negative, 1: positive). The dataset exhibits substantial class imbalance, with positive sample counts of 14,387, 589, and 1,838 for the respective tasks. We implement a stratified sampling method with a 4:1 ratio to divide the dataset into training and testing sets (D_{train}, D_{test}), ensuring consistent class distributions across both sets. To address the challenges of extreme class imbalance, we create a balanced training subset ($D_{\text{B-train}}$). This balanced subset contains an equal number of negative and positive samples sampled from the original D_{train}. The applications of these datasets are detailed in subsequent sections.

Implementation Details. In our experiments, WSIs are preprocessed by dividing them into multiple patches, each initially $1,024 \times 1,024$ pixels, and subsequently resized to 256×256 pixels for model input. The teacher model is trained on the balanced $D_{\text{B-train}}$, while the student model is trained on the comprehensive D_{train}, as detailed in section 2. The multi-task models are developed using PyTorch on an NVIDIA A100 GPU. We utilize the Adam optimizer with a fixed learning rate of 1×10^{-4} and a batch size of 16. A Cosine Annealing learning rate scheduler dynamically adjusted the learning rate throughout a 50-epoch training period for both teacher and student models. Two critical metrics for early screening tasks, the area under the receiver operating characteristic curve (AUC) and sensitivity (SEN), are reported in the subsequent experiments with the percentage form.

Table 1. Comparision with single-/multi-task models. The best results are **bolded**, and the second-best results are underlined.

Type	Model	Cancer		Candidiasis		Clue Cell		Average	
		AUC	SEN	AUC	SEN	AUC	SEN	AUC	SEN
Single-task	DT-free [22]	89.13	77.89	92.50	86.86	98.23	96.04	93.29	86.93
	TransMIL [23]	84.38	75.29	91.20	86.44	98.50	95.01	91.36	85.58
	LNPL-MIL [24]	89.96	85.12	92.31	89.37	98.43	**97.82**	93.57	90.77
	MambaMIL [25]	88.65	79.44	91.29	79.28	98.40	93.35	92.78	84.02
Multi-task	MOMA [26]	84.46	83.33	89.13	79.23	98.49	97.21	90.69	87.92
	MTDP [27]	77.01	73.61	85.23	78.81	92.28	86.36	84.84	79.59
	SALL [28]	81.85	81.12	86.62	64.40	98.07	96.18	88.85	80.57
	SSMTL-MD [29]	85.62	76.89	91.28	85.46	98.54	97.36	91.81	86.57
	MECDS (Teacher)	89.51	83.28	92.09	89.10	98.78	96.35	93.46	89.58
	MECDS (Student)	**90.42**	**87.14**	**93.04**	**90.09**	**99.08**	97.55	**94.18**	**91.59**

3.2 Comparison with Other Methods

To evaluate our proposed MECDS, we compare it with several state-of-the-art (SOTA) methods across three common TCT tasks: cervical cancer screening,

candidiasis detection, and clue cell diagnosis. Given the inherent challenge of low sensitivity towards positive samples when trained on imbalanced datasets, all compared methods are trained on the balanced $D_{\text{B-train}}$ and subsequently tested on the imbalanced D_{test} to assess their performance in realistic screening scenarios.

In the single-task setting, we compare our method with four robust single-task models: DT-free [22] (designed for cervical cancer screening), TransMIL [23], LNPL-MIL [24], and MambaMIL [25]. All models are independently trained for each specific task. In the multi-task setting, we select four medical multi-task models for comparison: MOMA [26], MTDP [27], SALL [28], and SSMTL-MD [29]. As illustrated in Table 1, even without employing the AKD training strategy, MECDS shows superior performance, outperforming existing multi-task methods and nearly matching the best results achieved by single-task models. Notably, all single-task models require separate training for each individual task, leading to increased computational overhead and higher storage requirements. The model's exceptional performance is directly attributed to its innovative dynamic feature routing strategy, which effectively mitigates inter-task interference while preserving the unique characteristics of individual tasks. Moreover, the incorporation of AKD further improved the model's performance, achieving SOTA results across all tasks.

3.3 Ablation study

We conduct ablation studies to evaluate the effectiveness of the proposed components (DFRM and AKD), as illustrated in Table 2. The results demonstrate the significant impact of the introduced DFRM across all tasks. Specifically, in cervical cancer screening, DFRM achieved remarkable improvements, yielding 2.8% and 7.96% gains in AUC and sensitivity, respectively, compared to the baseline model. These findings validate DFRM's capability to effectively select and transfer task-relevant features to the corresponding expert, thereby substantially enhancing predictive performance on WSIs. In terms of distillation, we find that directly applying conventional Kullback-Leibler (KL) [30] loss-based distillation yields only marginal improvements in AUC while significantly diminishing the model's ability to predict positive cases, as evidenced by a decline in sensitivity. In contrast, the proposed AKD loss circumvents this issue and even outperforms the teacher model. This superior performance is attributed to the AKD's focus on hard samples, which enables the student model to better assimilate the teacher model's knowledge, especially for positive cases.

3.4 Evaluation of Task Extensibility

The model's ability to extend to new tasks is a key capability of a multi-task model designed for early screening, underscoring its robustness and clinical applicability. To evaluate this, we conduct task extensibility experiments. We pretrain the model using any two of the three tasks as base tasks and then fine-tun it on the remaining task as an extension, to assess whether the model could

Table 2. Ablation study of the key components of our framework:(a) Dynamic Feature Routing Module (DFRM), (b) conventional KL loss-based distillation (KL) and (c) the proposed Asymmetric Knowledge Distillation (AKD).

Configuration			Cancer		Candidiasis		Clue Cell		Average	
DFRM	KL	AKD	AUC	SEN	AUC	SEN	AUC	SEN	AUC	SEN
×	×	×	86.71	75.32	88.96	84.29	98.14	95.55	91.27	85.05
✓	×	×	89.51	83.28	92.09	89.10	98.78	96.35	93.46	89.58
✓	✓	×	90.20	82.47	92.57	85.45	98.88	95.73	93.88	87.88
✓	×	✓	**90.42**	**87.14**	**93.04**	**90.09**	**99.08**	**97.55**	**94.18**	**91.59**

maintain performance across all three tasks. Specifically, for a trained MECDS, when a new task is introduced, one can simply add a new task token, a new task head, and a new router with a new expert in each MFA module, and only train these components on the new task's dataset, thereby facilitating computationally efficient learning of the new task. As shown in Table 3, the results indicate that our model can expand to new tasks while preserving nearly unchanged performance on previously learned tasks, achieving results comparable to those obtained when training on all tasks simultaneously. This demonstrates the model's excellent extensibility to new tasks, aligning well with the demands of real clinical scenarios.

Table 3. The performance of the task scalability experiments. "✓" indicates tasks selected for training, and "∗" represents tasks not trained during fine-tuning but used for multi-task testing in the final model.

Configuration			Cancer		Candidiasis		Clue Cell	
Cancer	Candidasis	Clue Cell	AUC	SEN	AUC	SEN	AUC	SEN
✓	✓	-	90.29	87.23	92.46	87.62	-	-
∗	∗	✓	89.60	87.11	91.73	87.62	98.73	97.35
✓	-	✓	90.92	87.18	-	-	99.03	97.29
∗	✓	∗	90.12	87.11	91.22	91.08	99.03	96.77
-	✓	✓	-	-	92.78	88.11	99.04	98.06
✓	∗	∗	90.18	86.21	95.65	87.98	98.84	97.16

4 Conclusion

In this study, we present MECDS, a unified framework for multi-task early cervical disease screening. The framework adapts a novel Multi-task Feature Adaptation strategy with Dynamic Feature Routing to effectively address inter-task interference by selectively processing task-relevant features. The Asymmetric Knowledge Distillation scheme successfully tackles the extreme class imbalance inherent in screening data with asymmetric focus levels for different samples.

Experiments demonstrate superior performance over specialized single-task and existing multi-task models, with excellent extensibility for accommodating new diagnostic tasks without comprehensive retraining. Future work will expand MECDS to cover a more comprehensive range of early cervical screening tasks.

Acknowledgments. This work was partially supported by AI4S Initiative and HPC Platform of ShanghaiTech University.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., Jemal, A.: Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians **68**(6), 394–424 (2018)
2. Gultekin, M., Ramirez, P.T., Broutet, N., Hutubessy, R.: World health organization call for action to eliminate cervical cancer globally. International Journal of Gynecological Cancer **30**(4), 426–427 (2020)
3. Siegel, R.L., Miller, K.D., Wagle, N.S., Jemal, A.: Cancer statistics, 2023. CA: A Cancer Journal for Clinicians **73**(1), 17–48 (2023)
4. Burd, E.M.: Human papillomavirus and cervical cancer. Clinical Microbiology Reviews **16**(1), 1–17 (2003)
5. Okunade, K.S.: Human papillomavirus and cervical cancer. Journal of Obstetrics and Gynaecology **40**(5), 602–608 (2020)
6. Koss, L.G.: The papanicolaou test for cervical cancer detection. a triumph and a tragedy. JAMA **261**(5), 737–743 (1989)
7. Landy, R., Pesola, F., Castañón, A., Sasieni, P.: Impact of cervical screening on cervical cancer mortality: Estimation using stage-specific results from a nested case–control study. British Journal of Cancer **115**(9), 1140–1146 (2016)
8. Godoy-Vitorino, F., Romaguera, J., Zhao, C., Vargas-Robles, D., Ortiz-Morales, G., Vázquez-Sánchez, F., Sanchez-Vázquez, M., de la Garza-Casillas, M., Martinez-Ferrer, M., White, J.R., et al.: Cervicovaginal fungi and bacteria associated with cervical intraepithelial neoplasia and high-risk human papillomavirus infections in a hispanic population. Frontiers in Microbiology **9**, 2533 (2018)
9. Zhang, X., Ji, J., Zhang, Q., Zheng, X., Ge, K., Hua, M., Cao, L., Wang, L.: A large annotated cervical cytology images dataset for ai models to aid cervical cancer screening. Scientific Data **12**(1), 23 (2025)
10. Fei, M., Shen, Z., Song, Z., Wang, X., Cao, M., Yao, L., Zhao, X., Wang, Q., Zhang, L.: Distillation of multi-class cervical lesion cell detection via synthesis-aided pretraining and patch-level feature alignment. Neural Networks **178**, 106405 (2024)
11. Cai, J., Xiong, H., Cao, M., Liu, L., Zhang, L., Wang, Q.: Progressive attention guidance for whole slide vulvovaginal candidiasis screening. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 233–242. Springer (2023)
12. Cheng, S., Liu, S., Yu, J., Rao, G., Xiao, Y., Han, W., Zhu, W., Lv, X., Li, N., Cai, J., et al.: Robust whole slide image analysis for cervical cancer screening using deep learning. Nature communications **12**(1), 5639 (2021)

13. Cao, M., Fei, M., Cai, J., Liu, L., Zhang, L., Wang, Q.: Detection-free pipeline for cervical cancer screening of whole slide images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 243–252. Springer (2023)
14. Wang, J., Yu, Y., Tan, Y., Wan, H., Zheng, N., He, Z., Mao, L., Ren, W., Chen, K., Lin, Z.: Artificial intelligence enables precision diagnosis of cervical cytology grades and cervical cancer. Nature Communications **15**(1), 4369 (2024)
15. Graham, B., El-Nouby, A., Touvron, H., Stock, P., Joulin, A., Jegou, H., Douze, M.: Levit: A vision transformer in convnet's clothing for faster inference. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12239–12249 (2021)
16. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. pp. 6000–6010. NIPS'17, Curran Associates Inc., Red Hook, NY, USA (2017)
17. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network (2015)
18. Yim, J., Joo, D., Bae, J., Kim, J.: A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7130–7138 (2017)
19. Zhou, B., Cui, Q., Wei, X.S., Chen, Z.M.: Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9716–9725 (2020)
20. Cao, K., Wei, C., Gaidon, A., Arechiga, N., Ma, T.: Learning imbalanced datasets with label-distribution-aware margin loss (2019)
21. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection (2018)
22. Cao, M., Fei, M., Cai, J., Liu, L., Zhang, L., Wang, Q., et al.: Detection-free pipeline for cervical cancer screening of whole slide images. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. pp. 243–252. Springer Nature Switzerland, Cham (2023)
23. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., Zhang, Y., et al.: Transmil: Transformer based correlated multiple instance learning for whole slide image classification. In: Advances in Neural Information Processing Systems. vol. 34, pp. 2136–2147. Curran Associates, Inc. (2021)
24. Shao, Z., Wang, Y., Chen, Y., Bian, H., Liu, S., Wang, H., Zhang, Y.: Lnpl-mil: Learning from noisy pseudo labels for promoting multiple instance learning in whole slide image. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 21438–21438 (2023)
25. Yang, S., Wang, Y., Chen, H.: MambaMIL: Enhancing Long Sequence Modeling with Sequence Reordering in Computational Pathology . In: proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. vol. LNCS 15004. Springer Nature Switzerland (October 2024)
26. Moon, S., Lee, H.: Moma: A multi-task attention learning algorithm for multi-omics data interpretation and classification. Bioinformatics **38**(8), 2287–2296 (2022)
27. Mormont, R., Geurts, P., Maree, R.: Multi-task pre-training of deep neural networks for digital pathology. IEEE Journal of Biomedical and Health Informatics **25**(2), 412–421 (2021)
28. Ju, L., Wang, X., Zhao, X., Lu, H., Mahapatra, D., Bonnington, P., Ge, Z.: Synergic adversarial label learning for grading retinal diseases via knowledge distillation and

multi-task learning. IEEE Journal of Biomedical and Health Informatics **25**(10), 3709–3720 (2021)
29. Gao, Z., Hong, B., Li, Y., Zhang, X., Wu, J., Wang, C., Zhang, X., Gong, T., Zheng, Y., Meng, D., et al.: A semi-supervised multi-task learning framework for cancer classification with weak annotation in whole-slide images. Medical Image Analysis **83**, 102652 (2023)
30. Kullback, S.: Kullback-leibler divergence (1951)