

# Medical-Knowledge Driven Multiple Instance Learning for Classifying Severe Abdominal Anomalies on Prenatal Ultrasound

Huanwen Liang<sup>1</sup>, Jingxian Xu<sup>1</sup>, Yuanji Zhang<sup>1</sup>, Yuhao Huang<sup>1</sup>, Yuhan Zhang<sup>1</sup>, Xin Yang<sup>1</sup>, Ran Li<sup>2</sup>, Xuedong Deng<sup>3</sup>, Yanjun Liu<sup>4</sup>, Guowei Tao<sup>5</sup>, Yun Wu<sup>6</sup>, Sheng Zhao<sup>7</sup>, Xinru Gao<sup>8</sup>, and Dong Ni<sup>1,9,10,11</sup>

<sup>1</sup> Medical Ultrasound Image Computing (MUSIC) Lab, School of Biomedical Engineering, Medical School, Shenzhen University, Shenzhen, China

<sup>2</sup> Shenzhen RayShape Medical Technology Co., Ltd, Shenzhen, China

<sup>3</sup> Center for Medical Ultrasound, The Affiliated Suzhou Hospital of Nanjing Medical University, Suzhou Municipal Hospital, Gusu School, Nanjing Medical University, Suzhou, China

<sup>4</sup> The First Hospital of China Medical University, Shenyang, China

<sup>5</sup> Qilu Hospital of Shandong University, Jinan, China

<sup>6</sup> Nanjing Women and Children's Healthcare Hospital, Women's Hospital of Nanjing Medical University, Nanjing, China

<sup>7</sup> Maternal and Child Health Hospital of Hubei Province, Wuhan, China

<sup>8</sup> Northwest Women's and Children's Hospital, Xian, China

<sup>9</sup> School of Artificial Intelligence, Shenzhen University, Shenzhen, China

<sup>10</sup> National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University, Shenzhen, China

<sup>11</sup> School of Biomedical Engineering and Informatics, Nanjing Medical University, Nanjing, China

nidong@szu.edu.cn; gxr\_ca@126.com

**Abstract.** Fetal abdominal malformations are serious congenital anomalies that require accurate diagnosis to guide pregnancy management and reduce mortality. Although AI has demonstrated significant potential in medical diagnosis, its application to prenatal abdominal anomalies remains limited. Most existing studies focus on image-level classification and rely on standard plane localization, placing less emphasis on case-level diagnosis. In this paper, we develop a case-level multiple instance learning (MIL)-based method, free of standard plane localization, for classifying fetal abdominal anomalies in prenatal ultrasound. Our contribution is three-fold. First, we adopt a mixture-of-attention-experts module (MoAE) to weight different attention heads for various planes. Secondly, we propose a medical-knowledge-driven feature selection module (MFS) to align image features with medical knowledge, performing self-supervised image token selection at the case-level. Finally, we propose a prompt-based prototype learning (PPL) to enhance the MFS. Extensively validated on a large prenatal abdominal ultrasound dataset containing 2,419 cases, with a total of 24,748 images and 6 categories, our proposed method outperforms the state-of-the-art competitors. Codes are available at: <https://github.com/LL-AC/AAcls>.

**Keywords:** Prenatal Abdominal Ultrasound · Multi-Instance Learning  
· Anomalies Classification

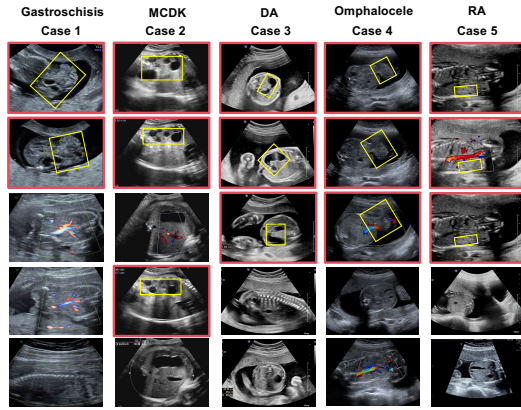
## 1 Introduction

Fetal abdominal malformations encompass a range of significant congenital anomalies that affect fetal health and are often associated with syndromes or aneuploidy [15]. The most severe and common abdominal anomalies include abdominal wall defects (such as omphalocele and gastroschisis), renal agenesis (RA), duodenal atresia (DA), and multicystic dysplastic kidney (MCDK), with an incidence of 1-4 per 10,000 live births [26,18,9]. Accurate diagnosis is critical for pregnancy decisions and treatment plans to reduce fetal morbidity and mortality. Ultrasound (US) remains the standard tool for prenatal abdominal evaluation, used to acquire standard planes and assess anatomical structures for diagnosing abdominal anomalies. However, abdominal anomalies are diverse in type and presentation, often occurring at earlier gestational ages. The detection of these anomalies heavily relies on the expertise and experience of sonographers, leading to varying detection rates and underscoring the ongoing challenges in prenatal diagnosis. To the best of our knowledge, artificial intelligence has demonstrated strong proficiency in diagnosing complex medical conditions [22,4,14]. However, research on its application to the diagnosis of prenatal abdominal anomalies remains limited.

**Fetal Anomalies Detection.** Currently, most studies for fetal anomalies detection primarily involve: (1) localizing standard planes from unordered *case-level* ultrasound images or video sequences and (2) detecting anomalies on the located *image-level* standard planes. Xie et al. [25] classified the fetal brain standard planes, and then the lesions in anomaly images can be localized by activation map. Lin et al. [12] developed an intelligent system, using the YOLO3 to detect fetal brain standard planes and brain structure to classify abnormal images. Qi et al. [16] constructed a deep learning model for nervous system anomalies classification in a multi-center fetal brain dataset. Most recently, Ciobanu et al. [2] proposed the first study to categorize fetal abdominal planes, achieving precise subsequent biomarker measurement and anomaly detection.

Although effective, the above studies require standard plane localization, which necessitates fine-grained image-level annotations. Meanwhile, they face challenges in cases such as abdominal wall defects (e.g., gastroschisis and omphalocele, shown in Fig. 1), where organ displacement or external exposure leads to the loss of abdominal structures, resulting in the failure to locate standard planes. To address these issues, we propose a Multiple Instance Learning (MIL)-based method that only requires case-level annotations (coarse-grained annotations from an ultrasound image pool taken from a single examination) and does not rely on localizing standard planes.

**Multiple Instance Learning (MIL)** represents a powerful tool for weakly supervised classification tasks. Specifically, it aims to aggregate multiple instances (image-level) into bag-level labels (case-level) [7,11,6,19,10,13,30,27,5].



**Fig. 1. Dataset of Multi-Center Prenatal Abdominal Ultrasound.** Each case contains several ultrasound images taken during a single examination, including both abnormal and normal images. Abnormal images are annotated with red boxes. Abnormal structures are annotated with yellow boxes.

Zhuang et al. [31] proposed a multi-scale feature fusion-based attention MIL model for case-level classification of thyroid nodules. Kaito et al. [20] proposed a selective aggregated MIL for case-level severity in ulcerative colitis diagnosis. Wang et al. [23] proposed a deep semi-supervised MIL framework for classifying diabetic macular edema from optical coherence tomography images.

The goal of this paper is to develop an MIL-based method that does not rely on localizing standard planes for case-level prenatal abdominal ultrasound anomaly classification. Our contribution is three-fold. (1) We adopt a mixture-of-attention-experts module (MoAE): By adopting MoAE, we can assign specific attention experts for different planes (i.e., sagittal, coronal, and transverse planes). (2) We propose a medical-knowledge-driven feature selection module: A self-supervised feature selection mechanism is designed based on medical knowledge, reducing the dependency on standard planes. (3) We propose a prompt-based prototype learning method to differentiate features across different categories. Experiments comparing our method with SOTA MIL methods show significant improvements in abdominal anomaly classification, providing a reliable and automated solution for prenatal ultrasound screening.

## 2 Methodology

Fig. 2 shows an overview of our proposed framework. First, a case image pool is fed into transformer layers for image integration. We adopt a mixture-of-attention-experts module (MoAE), which replaces the vanilla multi-head attention mechanism to control different attention experts, enabling feature integration across different planes. Secondly, a medical-knowledge-driven feature selection module (MFS) leverages medical knowledge to select significant images

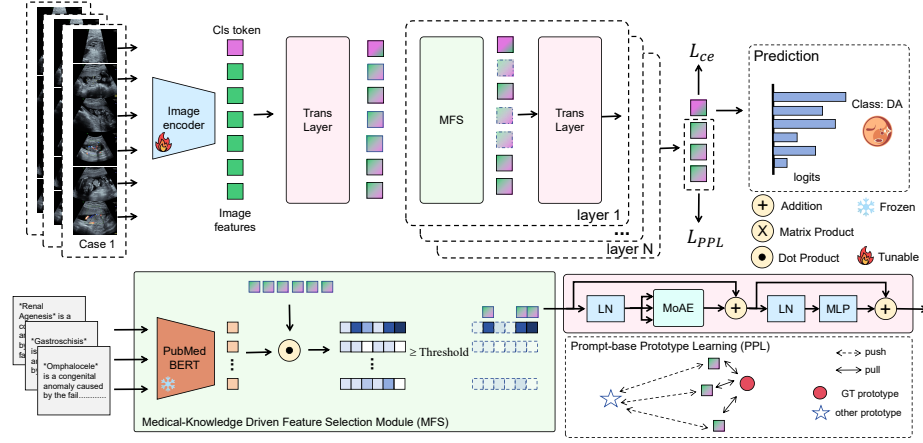


Fig. 2. Overview of our proposed framework

from the case image pool. Finally, the prompt-based prototype learning (PPL) module involves a loss function to bring paired medical knowledge and images closer in the feature space, ensuring the distinctiveness of the selected image.

**TransMIL Definition for Case-level Analysis.** The vanilla TransMIL [19] was designed for Whole Slide Imaging (WSI), including two main components: the image encoder and the transformer layers. Due to the large size, WSIs are usually divided into multiple non-overlapping patches as instances. The image encoder  $f$  first extracts features  $X'$  from these instances  $X$ , which are then fed into the transformer layers to generate the final prediction  $y$ . In our task, we treat a image pool  $X$  obtained from a single examination of a case as instances, each instance  $x_i$  represents one image, formulated by:

$$X' = f(X), X = (x_1, x_2, \dots, x_n), \quad (1)$$

$$y = \text{TransLayer}(h_{cls}, X'), X' = (x'_1, x'_2, \dots, x'_n). \quad (2)$$

**Mixture of Attention Experts Module for Integrating Planes.** For the transformer layer, multi-head attention is crucial, and different attention heads have different functions, including positional function [21]. Here, we consider different planes to possess spatial position information. We adopt the mixture-of-attention-experts module (MoAE) [8] as shown in Fig. 3, treating attention heads as experts within the MoE mechanism, enabling feature integration across different planes, formulated by:

$$\text{MoAE}(X') = \sum_{i=1}^h g_i H^i W_O^i, \quad (3)$$

$$H^i = \text{Attention}(X'W^Q, X'W^k, X'W^v), \quad (4)$$

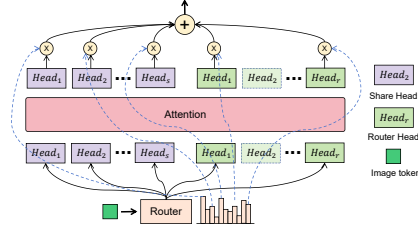


Fig. 3. Mixture of Attention Experts Module

$$g_i = \begin{cases} \alpha_1 \text{Softmax}(W_s x'_i)_j, & \text{if } 1 \leq j \leq h_s, \\ \alpha_2 \text{Softmax}(W_r x'_i)_j, & \text{if } (W_r x'_i)_j \in \text{Top-K}(\{(W_r x'_i)_j | h_s + 1 \leq j \leq h\}), \\ 0, & \text{otherwise.} \end{cases} \quad (5)$$

where  $H^i$  represents the output of the vanilla Attention mechanism.  $g_i$  represents the router score. Each token activates the top-K most relevant attention heads via routers. When the  $i_{th}$  attention head is activated,  $g_i > 0$ .  $\text{Softmax}(W_r x'_i)$  represents the router score.  $W$  represents a trainable projection matrix. This module also retains shared heads  $h_s$  to integrate common information across tokens. And the coefficients  $\alpha_1$  and  $\alpha_2$  balance the contributions of the shared and routed heads. This method enables each token to select the most relevant attention heads and produce a weighted summation based on router scores.

**Medical-Knowledge Driven Feature Selection Module.** Recent multi-modal pre-training models, such as CLIP [17], MedCLIP [24] and BioMedCLIP [29], have learned associations between images-text by training on large-scale language-image pairs using contrastive learning, aligning them in the feature space and enabling zero-shot or few-shot learning capabilities. Additionally, prompt learning has utilized a pre-trained Vision-Language Model (VLM) to adapt to new downstream tasks by incorporating task-specific prompts into the model’s input, demonstrating strong performance [1, 28].

Inspired by these, we propose a medical-knowledge-driven feature selection module (MFS). Guided by medical-knowledge prompts, The MFS performs self-supervised selection of abnormal images from the case-level image pool without relying on standard plane localization.

First, medical-knowledge are fed into a pre-trained model PubMedBERT [3] to extract the sentence feature as the *prompt*. The input format of the *prompts* is:  $Prompt = \{Cls, definition, signs\}$ . Secondly, we compute the similarity between each image token and the medical-knowledge *prompts* to obtain an activation score. Then, we use an adaptive threshold to filter out irrelevant images. The  $\beta$  increases progressively at each stage, the formula of *Threshold* can be written by:

$$Threshold = \beta * \frac{1}{n} \sum_{i=1}^n \max(\text{Softmax}(x'_i \cdot Prompts), \dim = C). \quad (6)$$

**Prompt-base Prototype learning for Enhancing MFS.** To ensure that the images selected by MFS are representative, we propose a loss function based on the image tokens  $X'$  as follows:

$$L_{PPL}(X', Prompts, Cls) = \sum \left( \frac{1}{n} \sum_{j=1}^n \frac{\exp(x'_j \cdot Prompts_{Cls})}{\sum_{i=1}^C \exp(x'_j \cdot Prompts_i)} \right), \quad (7)$$

where  $Prompts$  represents the set of all category prompts,  $C = \text{len}(Prompts)$  and  $Cls$  denotes the ground truth.  $N$  represents the size of the dataset, while  $n$  represents the number of images in a single case. The image token  $x'_j$  selected by MFS and the category prompt are used as a positive pair, while other prompts form negative pairs. Treating prompts as prototypes, we pull the image tokens of the positive pair closer to the prototype center and push them away from other centers, further enhancing the effectiveness of the selected features in MFS. We further add cross-entropy loss, and the overall loss function is:

$$L(y, Cls, Prompt) = L_{ce}(y, Cls) + L_{PPL}(X', Prompts, Cls). \quad (8)$$

### 3 Experimental Results

**Ethical Approval.** This study was approved by the Clinical Research Ethics Committee of the Affiliated Suzhou Hospital of Nanjing Medical University (K-2023-067-H01) and conducted following the Declaration of Helsinki. It was registered at [www.chictr.org.cn](http://www.chictr.org.cn) (ChiCTR2300071300).

**Materials and Implementation Details.** In this paper, we collected a multi-center prenatal abdominal ultrasound dataset. The age range of the pregnant women was 17 to 48 years, and the gestational age ranged from 10 to 39 weeks. Our dataset contains 2,419 cases and a total of 24,748 images, including abdominal wall defect and several common urological, digestive system anomalies: Duodenal Atresia (DA) (476), Gastroschisis (154), Omphalocele (760), Renal Agenesis (RA) (197), Multicystic Dysplastic Kidney (MCDK) (277), and Normal (555). Each case contains 1 to 65 images. The dataset was randomly split into training (1,516), validation (449), and test (454) sets at the case-level, ensuring that the distribution of categories was maintained across all subsets.

We implemented our method in PyTorch, using an NVIDIA 4090 GPU. The batch size, learning rate, and total epochs were set to 64, 1e-3, and 100, respectively. We used the Adam optimizer for model optimization. We used pre-trained model ResNet18 as image encoder. During training, we employed a warm-up learning rate strategy, starting with an initial learning rate of 1e-8, reaching the maximum learning rate at epoch 20, and then gradually decreasing using a half-cosine schedule to a minimum of 1e-8.

**Result.** Table 1 presents a comparison of our method with nine strong competitors, including Feature Max [6], Feature Mean [6], TransMIL [19], AbMIL [5], DSMIL [10], AdditiveMIL [7], IBMIL [13], Mamba2MIL [30], and SAMIL [20]. In anomalies classification, we primarily focus on the sensitivity for detecting abnormal fetuses. Therefore, we show sensitivity as the performance metric

**Table 1.** Method comparison. The best results are shown in bold.

	DA		Gastroschisis		Omphalocele		RA		MCDK		Normal		Acc
	Sen	F1	Sen	F1	Sen	F1	Sen	F1	Sen	F1	Sen	F1	
Features Max	87.50%	91.30%	46.88%	50.00%	82.89%	84.56%	75.00%	74.07%	82.14%	79.31%	<b>97.44%</b>	89.94%	78.64%
Features Mean	92.71%	92.71%	43.75%	46.67%	78.95%	81.91%	70.00%	74.67%	<b>87.50%</b>	82.35%	96.15%	88.76%	78.18%
TransMIL	93.75%	94.74%	53.12%	65.38%	<b>93.42%</b>	86.59%	67.50%	72.00%	73.21%	82.83%	96.15%	91.46%	79.53%
AbMIL	91.67%	89.34%	50.00%	57.14%	87.50%	<b>87.79%</b>	77.50%	<b>79.49%</b>	85.71%	88.89%	94.87%	89.16%	81.21%
DSMIL	<b>97.92%</b>	95.92%	46.88%	58.82%	90.79%	87.34%	72.50%	79.45%	85.71%	88.89%	93.59%	89.02%	81.23%
AdditiveMIL	90.62%	92.06%	50.00%	54.24%	83.55%	85.52%	67.50%	69.23%	<b>87.50%</b>	83.76%	94.87%	88.10%	79.01%
IBMIL	93.75%	92.78%	43.75%	53.85%	86.18%	85.62%	80.00%	72.73%	76.79%	81.90%	92.31%	88.34%	78.80%
Mamba2MIL	94.79%	96.3%	53.12%	56.67%	78.95%	83.33%	77.50%	61.39%	73.21%	81.19%	93.59%	86.39%	78.53%
SAMIL	92.71%	89.90%	40.62%	44.83%	79.61%	81.48%	80.00%	79.01%	80.36%	79.65%	87.18%	84.47%	76.75%
Our	<b>97.92%</b>	<b>97.92%</b>	<b>62.50%</b>	<b>67.80%</b>	87.50%	<b>87.79%</b>	<b>85.00%</b>	73.91%	80.36%	<b>89.11%</b>	<b>97.44%</b>	<b>94.41%</b>	<b>85.12%</b>

**Table 2.** Ablation studies. The best results are shown in bold.

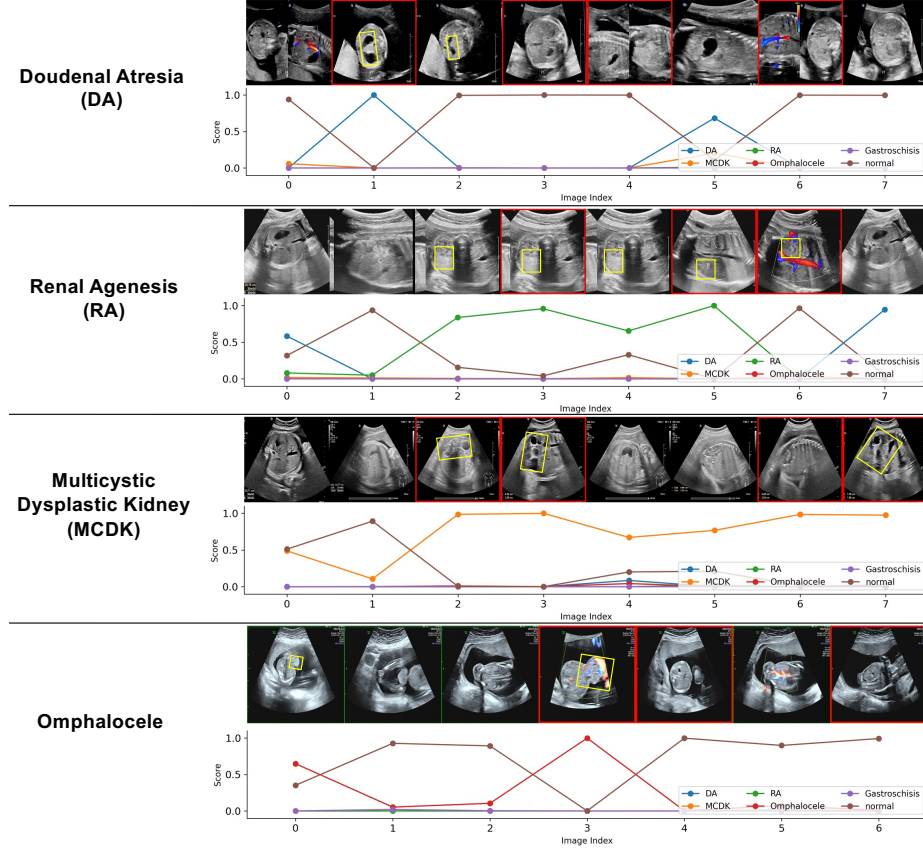
MFS MoAE PPL	DA		Gastroschisis		Omphalocele		RA		MCDK		Normal		Acc
	Sen	F1	Sen	F1	Sen	F1	Sen	F1	Sen	F1	Sen	F1	
✗ ✗ ✗	93.75%	94.74%	53.12%	65.38%	<b>93.42%</b>	86.59%	67.50%	72.00%	73.21%	82.83%	96.15%	91.46%	79.53%
✓ ✗ ✗	96.88%	97.38%	56.25%	58.06%	82.89%	84.85%	72.50%	69.05%	<b>85.71%</b>	88.89%	96.15%	90.36%	81.73%
✓ ✗ ✓	95.83%	96.34%	<b>62.50%</b>	59.70%	85.53%	86.09%	75.00%	<b>78.95%</b>	83.93%	87.85%	<b>1.00%</b>	<b>94.55%</b>	83.80%
✓ ✓ ✓	<b>97.92%</b>	<b>97.92%</b>	<b>62.50%</b>	<b>67.80%</b>	87.50%	<b>87.79%</b>	<b>85.00%</b>	73.91%	80.36%	<b>89.11%</b>	97.44%	94.41%	<b>85.12%</b>

for each category, while also presenting the F1 score. Further, we use weighted accuracy as the overall performance metric to account for class imbalance in the dataset. Results show that our method outperforms all other methods in most categories. Regarding sensitivity improvement, compared to the second-best model, the improvements are DA (3.13%), Gastroschisis (9.38%), RA (5%), and Normal (1.29%), respectively. In terms of F1 score, the improvements are DA (1.62%), Gastroschisis (2.42%), Omphalocele (0.45%), MCDK (0.22%), and Normal (2.95%), respectively. Additionally, our method achieves the best overall accuracy 85.12% among all the methods, demonstrating its effectiveness.

We also preformed ablation studies to evaluate the contribution of each proposed module in Table 2. Specifically, we treated the TransMIL as the baseline and added our proposed modules. When using the MFS module alone, it can be seen that MFS contributes about 2% improvement to the overall accuracy. Furthermore, the combination of MFS and PPL further enhances performance, resulting in an additional 2% improvement. As the modules were progressively added, the overall accuracy increased significant of 5.59%, highlighting the effectiveness of the proposed modules in enhancing model performance.

Fig. 4 displays typical cases using our modules. The images annotated with red boxes indicate those retained after applying our MFS, while the rest are discarded. As the number of image tokens decreases, the model becomes more focused on anomaly classification. The line chart shows that the images activated by MFS generally receive higher scores. This demonstrates that even if the model does not select all abnormal images, the chosen ones remain highly representative





**Fig. 4. Visualization of typical cases.** Abnormal structures are annotated with yellow boxes. The images activated by MFS are annotated with red boxes. The line chart represents the activation scores of the corresponding image categories.

as shown in Fig. 4 (RA and Omphalocele). These typical cases demonstrate that our method can perform case-level predictions effectively without relying on standard plane localization.

## 4 Conclusion

In this paper, we propose a medical-knowledge driven MIL method for improving case-level prenatal abdominal ultrasound anomalies classification. We first adopt MoAE, assigning different attention experts to specific planes enabling feature integration. Next, we propose the MFS, which enables the model to perform self-supervised selection of significant images from case-level ultrasound image pool based on medical knowledge. Finally, we propose a loss function PPL to enhance the MFS. Extensive experiments show the effectiveness of our method. In the



future, we will extend this framework to more congenital anomalies classification in fetus.

**Acknowledgments.** This work was supported by the grant from National Natural Science Foundation of China (12326619, 62171290); Science and Technology Planning Project of Guangdong Province (2023A0505020002); Frontier Technology Development Program of Jiangsu Province (BF2024078); Science and Technology Research and Development Program of Shaanxi Province (2011K12-05-11); Key Research and Development Program of Shaanxi (2025SF-YBXM-156).

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

## References

1. Chen, W., Si, C., et al.: Semantic prompt for few-shot image recognition. In: Proceedings of the IEEE/CVF Conference on CVPR. pp. 23581–23591 (2023)
2. Ciobanu, S.G., Enache, I.A., Iovoaica-Rănescu, C., Berbecaru, E.I.A., et al.: Automatic identification of fetal abdominal planes from ultrasound images based on deep learning. *Journal of Imaging Informatics in Medicine* pp. 1–8 (2025)
3. Gu, Y., Tinn, R., Cheng, H., Lucas, M., et al.: Domain-specific language model pre-training for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)* **3**(1), 1–23 (2021)
4. Huang, Y., Yang, X., Huang, X., Zhou, X., Chi, H., Dou, H., Hu, X., Wang, J., Deng, X., Ni, D.: Fourier test-time adaptation with multi-level consistency for robust classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 221–231. Springer (2023)
5. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning. pp. 2127–2136. PMLR (2018)
6. Ilse, M., Tomczak, J.M., Welling, M.: Deep multiple instance learning for digital histopathology. In: Handbook of Medical Image Computing and Computer Assisted Intervention, pp. 521–546. Elsevier (2020)
7. Javed, S.A., Juyal, D., Padigela, H., Taylor-Weiner, A., Yu, L., Prakash, A.: Additive mil: Intrinsically interpretable multiple instance learning for pathology. *Advances in Neural Information Processing Systems* **35**, 20689–20702 (2022)
8. Jin, P., Zhu, B., Yuan, L., Yan, S.: Moh: Multi-head attention as mixture-of-head attention. *arXiv preprint arXiv:2410.11842* (2024)
9. Laurichesse Delmas, H., et al.: Congenital unilateral renal agenesis: prevalence, prenatal diagnosis, associated anomalies. data from two birth-defect registries. *Birth defects research* **109**(15), 1204–1211 (2017)
10. Li, B., et al.: Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In: Proceedings of the IEEE/CVF conference on CVPR. pp. 14318–14328 (2021)
11. Li, H., Yang, F., et al.: Dt-mil: deformable transformer for multi-instance learning on histopathological image. In: Medical Image Computing and Computer Assisted Intervention: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24. pp. 206–216. Springer (2021)

12. Lin, M., He, X., Guo, H., He, M., Zhang, L., Xian, J., Lei, T., Xu, Q., Zheng, J., Feng, J., et al.: Use of real-time artificial intelligence in detection of abnormal image patterns in standard sonographic reference planes in screening for fetal intracranial malformations. *Ultrasound in Obstetrics & Gynecology* **59**(3), 304–316 (2022)
13. Lin, T., Yu, Z., Hu, H., Xu, Y., Chen, C.W.: Interventional bag multi-instance learning on whole-slide pathological images. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 19830–19839 (2023)
14. Liu, Z., Zhu, X., Han, T., Huang, Y., Wang, J., Liu, L., Wang, F., Ni, D., Gou, Z., Yang, X.: Mitral regurgitation recognition based on unsupervised out-of-distribution detection with residual diffusion amplification. In: *International Workshop on Machine Learning in Medical Imaging*. pp. 52–62. Springer (2024)
15. Pechriggl, E., Blumer, M., et al.: Embryology of the abdominal wall and associated malformations—a review. *Frontiers in Surgery* **9**, 891896 (2022)
16. Qi, Y., Cai, J., Lu, J., Xiong, R., Chen, R., Zheng, L., Ma, D.: Multi-center study on deep learning-assisted detection and classification of fetal central nervous system anomalies using ultrasound imaging. *arXiv preprint arXiv:2501.02000* (2025)
17. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., et al.: Learning transferable visual models from natural language supervision. In: *International conference on machine learning*. pp. 8748–8763. PMLR (2021)
18. Sarhan, O., Alghanbar, M., Alsulaim, A., et al.: Multicystic dysplastic kidney: impact of imaging modality selection on the initial management and prognosis. *Journal of pediatric urology* **10**(4), 645–649 (2014)
19. 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)
20. Shiku, K., Nishimura, K., Suehiro, D., Tanaka, K., Bise, R.: Ordinal multiple-instance learning for ulcerative colitis severity estimation with selective aggregated transformer. *arXiv preprint arXiv:2411.14750* (2024)
21. Voita, E., Talbot, D., Moiseev, F., Sennrich, R., Titov, I.: Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. pp. 5797–5808 (2019)
22. Wang, J., Miao, J., Yang, X., Li, R., Zhou, G., Huang, Y., Lin, Z., Xue, W., Jia, X., Zhou, J., et al.: Auto-weighting for breast cancer classification in multimodal ultrasound. In: *Medical Image Computing and Computer Assisted Intervention—MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VI* 23. pp. 190–199. Springer (2020)
23. Wang, X., Tang, F., Chen, H., Cheung, C.Y., Heng, P.A.: Deep semi-supervised multiple instance learning with self-correction for dme classification from oct images. *Medical Image Analysis* **83**, 102673 (2023)
24. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163* (2022)
25. Xie, B., Lei, T., Wang, N., Cai, H., et al.: Computer-aided diagnosis for fetal brain ultrasound images using deep convolutional neural networks. *International journal of computer assisted radiology and surgery* **15**, 1303–1312 (2020)
26. Xie, X., et al.: Global birth prevalence of major congenital anomalies: a systematic review and meta-analysis. *BMC Public Health* **25**, 449 (2025)
27. Yang, S., et al.: Mambamil: Enhancing long sequence modeling with sequence re-ordering in computational pathology. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 296–306. Springer (2024)

28. Yao, H., Zhang, R., Xu, C.: Visual-language prompt tuning with knowledge-guided context optimization. In: Proceedings of the IEEE/CVF conference on CVPR. pp. 6757–6767 (2023)
29. Zhang, S., Xu, Y., Usuyama, N., et al.: Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv preprint arXiv:2303.00915 (2023)
30. Zhang, Y., Zhang, X., Wang, J., Yang, Y., Peng, T., Tong, C.: Mamba2mil: State space duality based multiple instance learning for computational pathology. arXiv preprint arXiv:2408.15032 (2024)
31. Zhuang, L., Ivezic, V., et al.: Patient-level thyroid cancer classification using attention multiple instance learning on fused multi-scale ultrasound image features. In: AMIA Annual Symposium Proceedings. vol. 2023, p. 1344 (2024)