

AnomExpert: Identifying and Selecting Anatomical Planes for Prenatal Ultrasound Anomaly Diagnosis

Jian Wang^{1,2,3*}, Yang Yang^{4*}, Ziheng Pan⁵, Xiliang Zhu⁵, Yuhan Zhang⁴, Yanfeng Zhou^{1,3}, Dong Ni^{1,3,5}✉

¹Medical Ultrasound Image Computing (MUSIC) Lab, School of Artificial Intelligence, Shenzhen University, Shenzhen 518060, China

²College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China

³National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University, Shenzhen 518060, China

⁴School of Biomedical Engineering, Medical School, Shenzhen University, Shenzhen 518037, China

⁵School of Biomedical Engineering and Informatics, Nanjing Medical University, Nanjing 211166, China
nidong@szu.edu.cn

Abstract. Life-limiting congenital anomalies require accurate prenatal diagnosis for appropriate clinical decision-making. Prenatal ultrasound (US) examinations involve multiple anatomical planes, and diagnosis depends on identifying anatomical planes and selecting diagnostically relevant planes for each anomaly. Existing automated methods either rely on plane-level annotations or aggregate heterogeneous images without explicitly modeling these diagnostic capabilities. We propose **AnomExpert**, a prototype-driven framework for prenatal US anomaly diagnosis using only case-level supervision. AnomExpert introduces learnable plane prototypes to organize unordered images into latent representations corresponding to anatomical planes without requiring plane annotations. A disease-aware sparse selection mechanism further selects diagnostically relevant planes for each anomaly. Experiments on a multi-center dataset of 3,654 cases show that AnomExpert consistently outperforms nine representative multi-instance learning methods. Using a ViT-small backbone, it achieves 86.9% accuracy and 84.2% F1-score while maintaining parameter efficiency. These findings indicate that modeling anatomical plane identification and disease-specific plane selection improves weakly supervised multi-plane prenatal US anomaly classification. The code is available at <https://github.com/TianCat/AnomExpert>.

Keywords: Prenatal Ultrasound · Classification · Prototype Learning.

* Jian Wang and Yang Yang contributed equally to this work.

1 Introduction

Life-limiting congenital anomalies (LLCAs) are severe fetal conditions associated with high mortality [4]. Different anomalies require distinct management strategies. For example, severe open spina bifida may warrant prenatal fetal repair [1], complex congenital heart defects often require staged surgery after birth [15], and anencephaly is typically managed with palliative care or pregnancy termination [5]. Therefore, accurate prenatal differentiation among anomaly types is critical for appropriate clinical decision-making. Ultrasound (US) is the primary modality for screening and diagnosis [16]. In clinical practice, a typical prenatal US examination involves acquiring images from multiple anatomical planes. *Accurate diagnosis requires identifying the anatomical plane represented by each image and determining which planes provide diagnostically relevant evidence for a specific anomaly* (Fig. 1). However, this diagnostic process is time-consuming and cognitively demanding, as different anomalies rely on distinct subsets of anatomical planes. Moreover, US interpretation is inherently subjective, leading to variability in diagnostic consistency [17]. These challenges motivate automated methods for multi-plane prenatal diagnosis.

Developing automated systems that capture these diagnostic capabilities remains challenging. Existing approaches often follow staged pipelines, such as standard plane detection followed by anomaly classification [8,2]. While effective under curated conditions, these methods require fine-grained image-level annotations, which are costly and difficult to obtain. Moreover, they assume that standard anatomical planes can be reliably localized, an assumption often violated in clinical practice due to anatomical variability and abnormal presentations [6,9]. Multi-Instance Learning (MIL) provides a natural alternative by learning from unordered image sets using case-level supervision [3]. However, most MIL methods aggregate image features directly through attention or ranking mechanisms without explicitly modeling anatomical plane identity or disease-specific plane selection [10,11,18,13,12,20,19,14]. As a result, heterogeneous anatomical planes are mixed within a unified representation, limiting the model’s ability to capture plane-level semantics and disease-specific relevance.

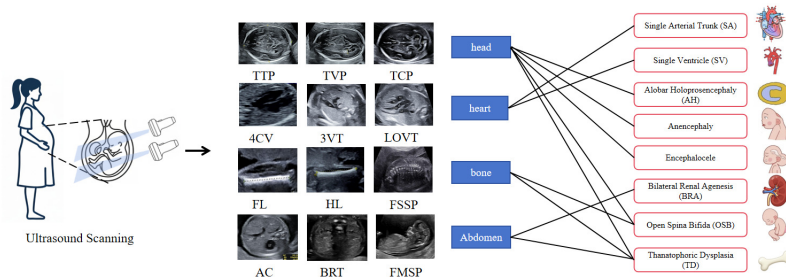


Fig. 1. Prenatal Ultrasound Anomaly Diagnosis

Motivated by the fundamental capabilities required for multi-plane diagnosis, we propose **AnomExpert**, a prototype-driven framework for prenatal US anomaly diagnosis using case-level supervision. Instead of replicating a staged pipeline, AnomExpert explicitly models two key capabilities: anatomical plane identification and disease-specific plane selection. Specifically, learnable plane prototypes organize unordered images into latent representations corresponding to anatomical planes without requiring plane annotations. A disease-aware sparse selection mechanism then adaptively identifies diagnostically relevant planes for each anomaly. This design enables the model to differentiate anatomical planes under case-level supervision and to selectively integrate disease-relevant planes for diagnosis. Our contributions are threefold: (1) a prototype-driven framework for weakly supervised anatomical plane identification without plane annotations; (2) a disease-aware sparse aggregation mechanism for adaptive diagnostic plane selection; and (3) extensive experiments demonstrating consistent improvements over representative MIL methods for prenatal US anomaly classification.

2 Methodology

As illustrated in Fig. 2, AnomExpert models two key capabilities of prenatal US diagnosis: anatomical plane identification and disease-specific plane selection. The framework first organizes unordered images into plane-level representations using learnable plane prototypes. These plane representations serve as candidate anatomical planes within each case. A disease-aware selection mechanism then identifies diagnostically relevant planes for anomaly prediction. The entire framework is trained end-to-end using only case-level labels.

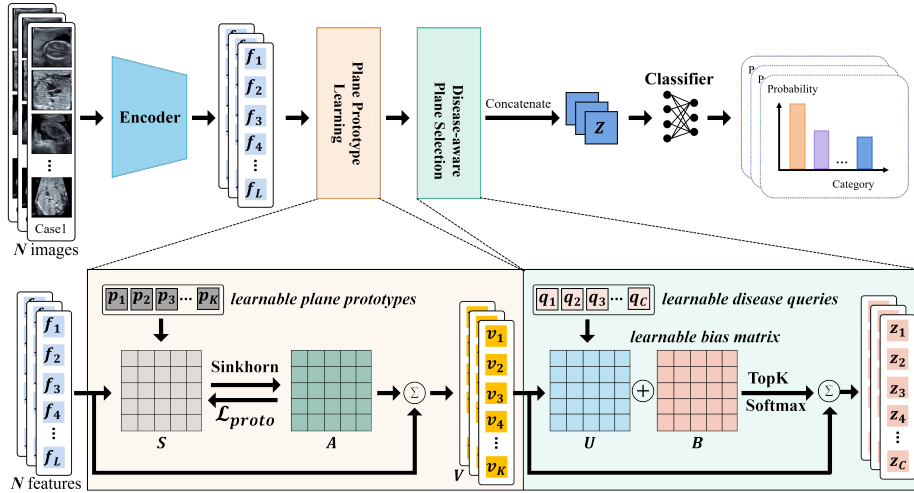


Fig. 2. Overview of the proposed AnomExpert

2.1 Plane Prototype Learning

A key challenge in multi-plane US diagnosis is identifying anatomical planes without explicit plane annotations. To approximate anatomical plane identification under case-level supervision, we introduce K learnable plane prototypes. Each prototype is intended to capture a distinct anatomical plane pattern. During training, image features are softly assigned to these prototypes, encouraging images with similar anatomical content to cluster around the same prototype.

Image embedding: Given an input image x_i , we extract a feature representation using a backbone network $\phi(\cdot)$ followed by a projection head $g(\cdot)$:

$$f_i = g(\phi(x_i)) \in \mathbb{R}^{256}. \quad (1)$$

Prototype assignment: We maintain a set of learnable plane prototypes $P = \{p_k \in \mathbb{R}^{256}\}_{k=1}^K$. For a mini-batch containing N images (summed across multiple cases), we compute cosine similarities between all image features and prototypes to form a score matrix $S \in \mathbb{R}^{N \times K}$. Soft assignments are obtained using Sinkhorn-balanced normalization [7]:

$$A = \text{Sinkhorn}(\exp(S/\epsilon_s)) \in \mathbb{R}^{N \times K}, \quad (2)$$

where $\epsilon_s = 0.05$ controls assignment sharpness. Sinkhorn balancing encourages diverse prototype utilization and helps prevent prototype collapse.

Prototype loss: To encourage consistent image-to-prototype assignments, we optimize a prototype assignment loss:

$$\mathcal{L}_{proto} = -\frac{1}{N} \sum_{i=1}^N \sum_{k=1}^K A_{i,k} \log \text{softmax}(S_{i,:}/\tau)_k, \quad (3)$$

where τ is a learnable temperature parameter. Through this learning process, images with similar anatomical characteristics are grouped under the same prototype, enabling anatomical plane identification without plane labels.

Plane representation: Given a case with L images, we aggregate image features into plane representations using the corresponding assignment weights:

$$v_k = \frac{\sum_{i=1}^L A_{i,k} f_i}{\sum_{i=1}^L A_{i,k} + \delta}, \quad k = 1, \dots, K, \quad (4)$$

where δ is a small constant for numerical stability. This yields a structured plane representation set $V = \{v_k \in \mathbb{R}^{256}\}_{k=1}^K$ for the current case, which serves as the basis for downstream disease-specific plane selection.

2.2 Disease-aware Plane Selection

Unlike conventional MIL approaches that aggregate features from all images, our method performs selection at the plane-level. To this end, we introduce learnable disease queries to estimate the relevance of each anatomical plane to each disease.

Disease queries: We define a set of learnable disease queries:

$$Q = \{q_c \in \mathbb{R}^{256}\}_{c=1}^C, \quad (5)$$

where C is the number of categories.

Relevance estimation: Given the structured plane representations $V = \{v_k \in \mathbb{R}^{256}\}_{k=1}^K$ obtained in Sec. 2.1, we compute cosine similarity scores between each disease query and all plane representations:

$$u_{c,k} = \text{sim}(q_c, v_k). \quad (6)$$

To capture disease-specific preferences over anatomical planes, we introduce a learnable bias matrix $B \in \mathbb{R}^{C \times K}$, which models prior associations between diseases and planes. The final relevance scores are obtained as:

$$\tilde{u}_{c,k} = u_{c,k} + B_{c,k}. \quad (7)$$

Sparse plane selection and aggregation: For each disease c , we retain the top- k most relevant planes:

$$\Omega_c = \text{TopK}(\tilde{u}_{c,:}), \quad (8)$$

where Ω_c denotes the selected plane indices. We then compute disease-specific representations by aggregating the selected planes using normalized weights:

$$z_c = \sum_{k \in \Omega_c} \alpha_{c,k} v_k, \quad \alpha_{c,\Omega_c} = \text{softmax}(\tilde{u}_{c,\Omega_c}/T), \quad (9)$$

where $T = 0.07$ is a temperature constant controlling the sharpness of selection. This mechanism ensures that anomaly prediction is based on a subset of diagnostically relevant anatomical planes rather than all available images.

Case-level prediction: The disease-specific representations are concatenated and fed into a classifier to obtain the final prediction:

$$\hat{y} = \text{Classifier}([z_1, \dots, z_C]). \quad (10)$$

Training objective: The model is trained end-to-end using case-level category label y . We use cross-entropy loss for anomaly classification:

$$\mathcal{L}_{cls} = \text{CE}(\hat{y}, y), \quad (11)$$

and jointly optimize the prototype assignment loss defined in Equation 3:

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda \mathcal{L}_{proto}, \quad (12)$$

where λ balances plane prototype learning and disease classification.

Table 1. Performance comparison of different methods on the test set (mean $\pm$ std over five runs). Param. denotes model parameters (M), and all metrics are in percentage (%). Best and second-best results are shown in blue and underlined, respectively.

Methods	Param.	Accuracy	Precision	Recall	F1-score	AUC
AttentionMIL	11.31	81.7 $\pm$ 1.08	80.0 $\pm$ 1.79	78.2 $\pm$ 1.24	78.4 $\pm$ 1.39	96.4 $\pm$ 0.47
MeanNetMIL	23.53	82.2 $\pm$ 0.80	80.9 $\pm$ 1.00	79.3 $\pm$ 1.50	79.3 $\pm$ 1.40	96.7 $\pm$ 0.30
MaxNetMIL	21.29	<u>85.5$\pm$1.06</u>	<u>84.0$\pm$1.52</u>	<u>81.5$\pm$1.47</u>	<u>82.1$\pm$1.59</u>	96.7 $\pm$ 0.32
TransMIL	30.91	83.5 $\pm$ 1.30	82.2 $\pm$ 2.00	81.0 $\pm$ 2.01	81.2 $\pm$ 2.00	96.9 $\pm$ 0.50
DSMIL	23.97	78.1 $\pm$ 2.54	77.2 $\pm$ 2.40	75.8 $\pm$ 2.84	75.0 $\pm$ 2.82	95.8 $\pm$ 0.60
AdditiveMIL	11.97	77.9 $\pm$ 0.91	74.7 $\pm$ 1.52	73.7 $\pm$ 1.31	74.0 $\pm$ 1.40	95.8 $\pm$ 0.34
Mamba2MIL	39.43	82.3 $\pm$ 4.49	82.1 $\pm$ 3.78	81.8 $\pm$ 3.99	81.2 $\pm$ 4.05	97.1 $\pm$ 0.97
SAMIL	27.72	84.1 $\pm$ 0.52	82.2 $\pm$ 1.49	80.0 $\pm$ 0.83	80.3 $\pm$ 1.05	96.4 $\pm$ 0.41
AAcls-MIL	19.07	83.6 $\pm$ 1.18	82.3 $\pm$ 1.28	79.8 $\pm$ 1.03	80.3 $\pm$ 0.86	97.3 $\pm$ 0.31
Anom. (ViT-t)	5.60	83.8 $\pm$ 0.96	81.8 $\pm$ 0.93	80.4 $\pm$ 1.39	80.6 $\pm$ 1.24	97.3 $\pm$ 0.23
Anom. (ViT-s)	21.80	<u>86.9$\pm$0.72</u>	<u>84.9$\pm$1.12</u>	<u>83.9$\pm$0.66</u>	<u>84.2$\pm$0.69</u>	<u>97.9$\pm$0.20</u>

3 Experiments and Results

3.1 Datasets and Experimental Setup

Datasets In this study, we collected a multi-center prenatal US dataset from 24 medical centers with approval from the institutional review board; informed consent was waived due to the retrospective study design. The dataset comprises 3,654 cases with 61,460 images, averaging 17 images per case (range: 5–94, std=13). Each case was assigned a single primary diagnosis. All diagnoses were confirmed by experienced fetal US specialists according to established clinical criteria. The dataset includes eight types of LLCAs and normal controls: Single Arterial Trunk (SA, n=399), Single Ventricle (SV, n=341), Alobar Holoprosencephaly (AH, n=478), Anencephaly (n=643), Encephalocele (n=180), Bilateral Renal Agenesis (BRA, n=335), Open Spina Bifida (OSB, n=348), Thanatophoric Dysplasia (TD, n=410), and Normal (n=520). Gestational ages ranged from 10 to 39 weeks (mean $\pm$ std: 20.7 $\pm$ 12.0 weeks). The dataset was randomly partitioned at the case level into training (n=2,558), validation (n=366), and testing (n=730) sets to prevent data leakage. The splits were stratified to preserve class distribution, and gestational age distributions were comparable across subsets.

Implementation Details We evaluated our method using two ImageNet-pretrained backbone architectures, ViT-tiny (ViT-t) and ViT-small (ViT-s), to assess robustness across different model capacities. The number of plane prototypes was set to $K = 30$, and the disease-aware selection retained the top- $k = 4$ planes for each category. The weight of the prototype assignment loss was fixed at $\lambda = 0.1$. Models were trained for 60 epochs using Adam (initial learning rate 1×10^{-4} , weight decay 1×10^{-5}), and the checkpoint with the best validation performance was used for testing. We used cosine annealing for learning rate decay and a batch size of 8 (case-level). Input images were randomly cropped and

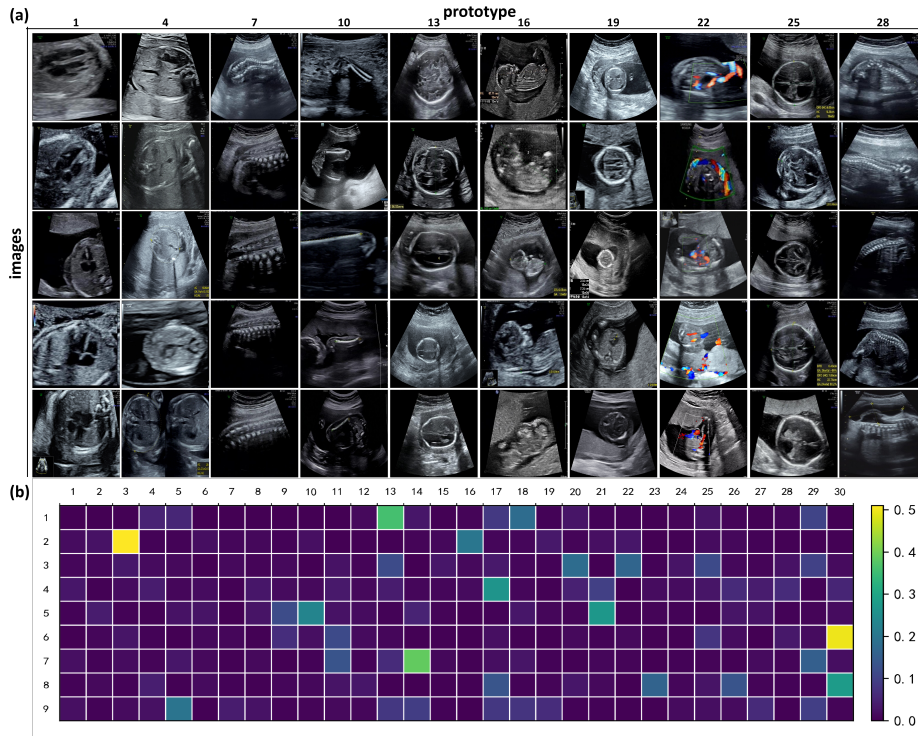


Fig. 3. Qualitative visualization. (a) Test images assigned to selected plane prototypes. (b) Class-wise relevance heatmap showing disease-plane associations.

resized to 224×224 . Additional data augmentations included horizontal flipping, color jittering, affine transformations, and random grayscale conversion.

Evaluation Protocol We compared our method with nine representative multi-instance learning (MIL) approaches, including AttentionMIL [10], MeanNetMIL [11], MaxNetMIL [11], TransMIL [18], DSMIL [13], AdditiveMIL [12], Mamba2MIL [20], SAMIL [19], and AAcls-MIL [14]. All competing methods were implemented using their official open-source code and evaluated under the same data split. Performance was evaluated using Accuracy, macro-averaged Precision, Recall, F1-score, and area under the ROC curve (AUC). Each method was trained and evaluated over five independent runs with different random seeds. We report the mean and standard deviation across runs.

3.2 Results

Quantitative Results Table 1 presents the quantitative comparison between AnomExpert and nine representative MIL methods on the test set. Overall, the

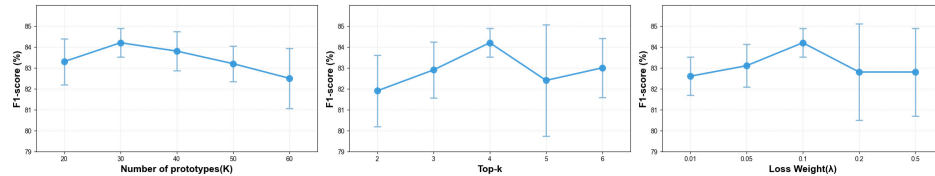
Table 2. Component ablations of AnomExpert on the test set (mean $\pm$ std over five runs). All metrics are reported in percentage (%).

Variant	Accuracy	F1-score
Full AnomExpert	86.9 $\pm$ 0.72	84.2 $\pm$ 0.69
Plane Prototype Learning		
w/o prototype loss ($\lambda = 0$)	84.8 $\pm$ 1.18	82.7 $\pm$ 1.43
Softmax assignment (w/o Sinkhorn balancing)	83.4 $\pm$ 0.72	80.8 $\pm$ 0.78
Disease-aware Plane Selection		
Dense aggregation (w/o top- k)	86.2 $\pm$ 1.81	83.7 $\pm$ 1.81
w/o disease bias ($B = 0$)	85.1 $\pm$ 1.22	82.4 $\pm$ 1.46

ViT-small variant of AnomExpert achieves the best performance across all evaluation metrics. Specifically, it achieves 86.9% accuracy, 84.9% precision, 83.9% recall, 84.2% F1-score, and 97.9% AUC. Compared with the best-performing baseline, MaxNetMIL, AnomExpert improves accuracy by 1.4 percentage points (86.9% vs. 85.5%) and F1-score by 2.1 percentage points (84.2% vs. 82.1%). Furthermore, AnomExpert demonstrates strong parameter efficiency. Using the ViT-tiny backbone, it achieves competitive performance (83.8% accuracy and 97.3% AUC) with only 5.60M parameters, fewer than all baseline models. These results confirm the effectiveness of AnomExpert in improving performance and parameter efficiency over existing MIL methods.

Qualitative Results Fig. 3(a) shows test images assigned to selected plane prototypes. Images grouped under the same prototype exhibit consistent anatomical patterns, indicating that the learned prototypes approximate anatomical plane identity without requiring plane annotations. Fig. 3(b) shows the class-wise plane selection heatmap. The heatmap reflects the average selection probabilities of planes across all test cases. Certain planes are selected more frequently, suggesting that they correspond to diagnostically informative anatomical planes. Together, these visualizations demonstrate that AnomExpert organizes images into plane representations and performs selective aggregation for prediction.

Ablation Study Table 2 shows that removing prototype learning ($\lambda = 0$) reduces F1-score by 1.5 points, while replacing Sinkhorn with softmax leads to a

**Fig. 4.** Ablation study on key hyperparameters. F1-score versus plane prototype number K (left), top- k selection (middle), and loss weight λ (right).

larger drop to 80.8%. For disease-aware selection, dense aggregation and removing the disease bias matrix both degrade performance. These results indicate that both plane identification and disease-aware selection contribute to overall performance. Fig. 4 illustrates the influence of key hyperparameters. We evaluated five candidate values for each hyperparameter, and the best performance was achieved with $K = 30$, top- $k = 4$, and $\lambda = 0.1$.

4 Conclusion

We proposed AnomExpert, a prototype-driven framework for prenatal US anomaly diagnosis under case-level supervision. By modeling anatomical plane identification and disease-specific plane selection, the method integrates key diagnostic capabilities into MIL. Experiments show consistent improvements over representative MIL methods, and ablation studies support the contribution of each component. These results highlight the importance of combining anatomical plane modeling and disease-aware plane selection for improving weakly supervised multi-plane medical image analysis. Future research could explore enhancing the scalability of AnomExpert for more diverse datasets and integrating it with clinical workflows for real-time anomaly detection and diagnosis.

Acknowledgments. This work is supported by the Frontier Technology Development Program of Jiangsu Province (No. BF2024078), National Natural Science Foundation of China (No. 12326619, 62572324), Science and Technology Planning Project of Guangdong Province (No. 2023A0505020002).

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

References

1. Adzick, N.S., Thom, E.A., Spong, C.Y., Brock III, J.W., Burrows, P.K., Johnson, M.P., Howell, L.J., Farrell, J.A., Dabrowiak, M.E., Sutton, L.N., et al.: A randomized trial of prenatal versus postnatal repair of myelomeningocele. *New England Journal of Medicine* **364**(11), 993–1004 (2011)
2. Arnaout, R., Curran, L., Zhao, Y., Levine, J.C., Chinn, E., Moon-Grady, A.J.: An ensemble of neural networks provides expert-level prenatal detection of complex congenital heart disease. *Nature medicine* **27**(5), 882–891 (2021)
3. Barbosa, D., Ferreira, M., Junior, G.B., Salgado, M., Cunha, A.: Multiple instance learning in medical images: a systematic review. *IEEE Access* **12**, 78409–78422 (2024)
4. Breeze, A.C., Lees, C.C.: Antenatal diagnosis and management of life-limiting conditions. In: *Seminars in Fetal and Neonatal Medicine*. vol. 18, pp. 68–75. Elsevier (2013)
5. Cook, R.J., Erdman, J.N., Hevia, M., Dickens, B.M.: Prenatal management of anencephaly. *International Journal of Gynecology & Obstetrics* **102**(3), 304–308 (2008)

6. Coronado-Gutierrez, D., Eixarch, E., Monterde, E., Matas, I., Traversi, P., Gratacos, E., Bonet-Carne, E., Burgos-Artizzu, X.P.: Automatic deep learning-based pipeline for automatic delineation and measurement of fetal brain structures in routine mid-trimester ultrasound images. *Fetal diagnosis and therapy* **50**(6), 480–490 (2023)
7. Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems* **26** (2013)
8. Guo, J., Tan, G., Wu, F., Wen, H., Li, K.: Fetal ultrasound standard plane detection with coarse-to-fine multi-task learning. *IEEE Journal of Biomedical and Health Informatics* **27**(10), 5023–5031 (2022)
9. Huang, Z., Wessler, B.S., Hughes, M.C.: Detecting heart disease from multi-view ultrasound images via supervised attention multiple instance learning. In: *Machine Learning for Healthcare Conference*. pp. 285–307. PMLR (2023)
10. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: *International conference on machine learning*. pp. 2127–2136. PMLR (2018)
11. 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)
12. 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)
13. 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)
14. Liang, H., Xu, J., Zhang, Y., Huang, Y., Zhang, Y., Yang, X., Li, R., Deng, X., Liu, Y., Tao, G., et al.: Medical-knowledge driven multiple instance learning for classifying severe abdominal anomalies on prenatal ultrasound. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 344–354. Springer (2025)
15. Ohye, R.G., Schranz, D., D’Udekem, Y.: Current therapy for hypoplastic left heart syndrome and related single ventricle lesions. *Circulation* **134**(17), 1265–1279 (2016)
16. Salomon, L., Alfirevic, Z., Berghella, V., Bilardo, C., Chalouhi, G., Costa, F.D.S., Hernandez-Andrade, E., Malinger, G., Munoz, H., Paladini, D., et al.: Isuog practice guidelines (updated): performance of the routine mid-trimester fetal ultrasound scan. *Ultrasound in Obstetrics and Gynecology* **59**(6), 840–856 (2022)
17. Sarris, I., Ioannou, C., Chamberlain, P., Ohuma, E., Roseman, F., Hoch, L., Altman, D., Papageorghiou, A., Fetal, I., for the 21st Century (INTERGROWTH-21st), N.G.C.: Intra-and interobserver variability in fetal ultrasound measurements. *Ultrasound in obstetrics & gynecology* **39**(3), 266–273 (2012)
18. 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)
19. Shiku, K., Nishimura, K., Suehiro, D., Tanaka, K., Bise, R.: Ordinal multiple-instance learning for ulcerative colitis severity estimation with selective aggregated transformer. In: *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. pp. 4290–4299. IEEE (2025)
20. 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)