GCM-Net: Anatomy-Aware Gaussian-Contrastive Multi-Layer Fusion Network for Abdominal Ultrasound Standard Plane Classification

Anqi Wang[1†], Zihan Li[1†], Lei Zhang[2], Xinye Zhang[1], Yingni Wang[1], Xinshan Yang[1], Mingxuan Liu[1], Xiuming Wang[2], Hongen Liao[1], and Guochen Ning[1✉]

[1] School of Biomedical Engineering, Tsinghua University, Beijing, P.R. China
ningguochen@tsinghua.edu.cn
[2] School of Clinical Medicine (Beijing Tsinghua Changgung Hospital), Beijing, P.R. China

Abstract. Reliable classification of standard abdominal ultrasound planes is essential for sonography and autonomous diagnosis. However, this task remains challenging due to the inherent low contrast and speckle noise in ultrasound imaging, further compounded by subtle inter-plane differences within subjects and substantial anatomical variations across the population. Here, we propose the Gaussian-contrastive multi-layer fusion network (GCM-Net), which comprises a learnable multi-layer feature fusion (LMFF) module to synergistically integrate hierarchical spatial details with high-level semantics, an adaptive anisotropic Gaussian refinement (AAGR) module to spatially focus on critical anatomical structures, and a supervised contrastive learning branch to enhance feature discriminability. Systematic evaluation on the publicly available FPUS23 dataset for fetal ultrasound and the private AbdoSP-9 dataset for abdominal ultrasound demonstrates that GCM-Net outperformed existing methods across key classification metrics. Additionally, GCM-Net demonstrated robust performance in real-time causal processing of clinical ultrasound videos, achieving an expert-validated accuracy of 81.25% through the temporal confidence refinement (TCR) framework. These results underscore GCM-Net's potential not only to streamline routine clinical diagnosis but also to serve as a high-level controller for autonomous robotic ultrasound. Code: https://github.com/diligentwaq-spec/GCM-Net

Keywords: Ultrasound Imaging · Standard Plane Classification · Feature Fusion · Contrastive Learning.

1 Introduction

Abdominal ultrasound is pivotal for clinical diagnosis due to its real-time, non-invasive nature [1, 2]. Its diagnostic validity relies on acquiring and interpreting standard planes (SPs), which are anatomical sections rich in morphological

[†]Equal contribution

and pathological information. These SPs underpin biometrics and disease assessment [3] and serve as critical visual landmarks for autonomous robotic navigation and standardized image acquisition [4–7]. However, recognizing SPs from continuous scanning sequences remains challenging due to low contrast, speckle noise, substantial anatomical variations across the population, and subtle differences between adjacent planes [8], leading to high operator dependency [9]. A robust, structured visual representation framework is therefore essential for reliable SP recognition, a critical step toward standardized clinical diagnosis and autonomous robotic ultrasound.

Deep learning has significantly advanced SP recognition. Early approaches relied on convolutional neural networks (CNNs) such as ResNet and VGG to extract localized structural features [10, 11], later improved by lightweight pyramid convolution (LPC) modules [12], yet remained constrained by limited receptive fields. To address this, Transformer-based architectures and hybrid frameworks integrating state-space models (SSMs) were introduced to leverage self-attention for modeling long-range dependencies [13–15].

Despite these architectural improvements, imaging artifacts and anatomical variations still induce severe visual ambiguity in SP recognition [16]. Moreover, most existing studies remain limited to single-organ settings and a few highly distinguishable planes, failing to meet the comprehensive screening requirements of clinical practice [17]. Consequently, the high intra-class variability and subtle inter-class differences of ultrasound images still cause elevated false-positive rates when distinguishing morphologically similar planes [18].

Effective visual representations are pivotal for accurate SP classification. Recent visual foundation models, especially those pre-trained on large-scale ultrasound datasets, demonstrate strong feature extraction capability [19, 20]. However, adapting these models directly to SP recognition remains challenging. Their generic representations often lack the task-specific granularity needed to distinguish subtle inter-plane differences, which are critical for clinical reliability.

To address this gap, we propose GCM-Net (Gaussian-Contrastive Multi-layer Fusion Network), a new anatomy-aware network that integrates cross-layer feature fusion with anatomical geometric constraints. The network leverages a pre-trained foundation model for strong general representation capacity and introduces a learnable multi-layer feature fusion (LMFF) module to adaptively aggregate multi-scale information. An adaptive anisotropic Gaussian refinement (AAGR) module further improves interpretability by generating spatial weight maps for focused enhancement in key regions. In addition, a memory-bank-based supervised contrastive learning strategy strengthens feature discriminability under low-contrast and highly similar imaging conditions.

We evaluated our method on a self-constructed abdominal ultrasound dataset containing nine categories of SPs, as well as on one public FPUS23 [21] dataset. The results demonstrate that GCM-Net outperforms existing methods across multiple metrics. Crucially, experimental validation in ultrasound video streams proves that GCM-Net achieves stable and high-confidence automated assessment, indicating strong potential for clinical application.

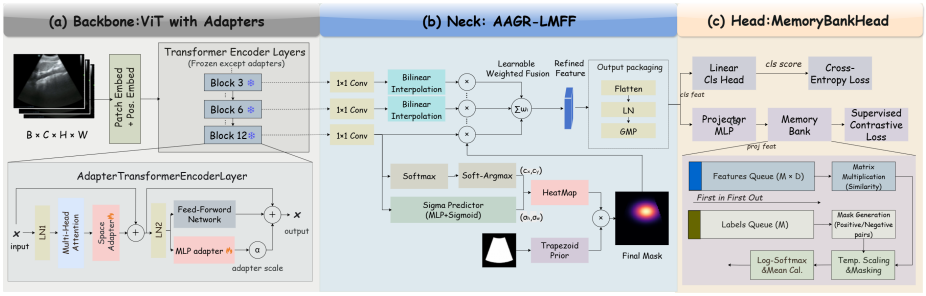

Fig. 1. Architecture of proposed GCM-Net. (a) Backbone: Vision Transformer (ViT) with adapters for parameter-efficient fine-tuning. (b) Neck: adaptive anisotropic Gaussian refinement (AAGR) and learnable multi-layer feature fusion (LMFF). (c) Head: dual-branch head consisting of a linear head for class-score prediction and a projection head for contrastive representation learning.

2 Methodology

GCM-Net achieves precise classification of SPs in continuous scanning videos by utilizing three key innovations (Fig. 1): (a) adaptation of a medical foundation model for robust feature representation, (b) multi-scale feature fusion and spatial refinement module for enhanced discrimination, and (c) dual-head optimization strategy employing memory-bank-augmented contrastive loss to enhance anatomical distinctiveness.

2.1 Feature Extraction via MED-SA.

GCM-Net adopts the MED-SA backbone [22] with Parameter-Efficient Fine-Tuning (PEFT) for robust and efficient image representations from ultrasound images. Specifically, the pre-trained ViT [19] is frozen with lightweight adapters inserted. For the l-th layer, let z_l denote the input feature representation. The intermediate state z'_l and the subsequent output z_{l+1} are formulated as:

$$z'_l = \text{Adap}_s(\text{MSA}(\text{LN}(z_l))) + z_l, z_{l+1} = \text{FFN}(\text{LN}(z'_l)) + z'_l + s \cdot \text{Adap}_m(\text{LN}(z'_l)) \quad (1)$$

where s denotes the scaling factor for the residual adaptation, LN represents Layer Normalization, and MSA is Multi-head Self-Attention. Here, Adap_s and Adap_m represent the spatial and MLP adapters respectively. The multi-scale feature pyramid $F = \{F_3, F_6, F_{12}\}$ is constructed by extracting and reshaping the outputs from specific layers, utilizing F_{12} as the primary semantic anchor.

2.2 Learnable Multi-scale Feature Fusion (LMFF)

To effectively aggregate local textural details and global semantic contexts from the learned representations, we propose the learnable multi-scale feature fusion

(LMFF) module. Unlike static summation or concatenation, LMFF introduces a set of learnable scalar weights $\alpha = \{\alpha_3, \alpha_6, \alpha_{12}\}$ to dynamically re-weight features based on their contribution to the recognition task. The weights are normalized via *Softmax*. By weighted summation of spatially aligned maps, the fused feature F_{fuse} is obtained:

$$\omega_l = \frac{\exp(\alpha_l)}{\sum_{k\in\{3,6,12\}} \exp(\alpha_k)}, \quad l \in \{3, 6, 12\} \tag{2}$$

$$F_{\text{fuse}} = \sum_{l\in\{3,6,12\}} \omega_l \cdot \psi(F_l) \tag{3}$$

where $\psi(\cdot)$ denotes the transformation consisting of a 1×1 lateral convolution for channel unification followed by bilinear interpolation to align all feature maps to the spatial resolution of the master feature F_{12}.

2.3 Adaptive Anisotropic Gaussian Refinement (AAGR)

To effectively focus the model's attention on the diagnostic Region of Interest amidst inherent speckle noise and background artifacts, we propose the adaptive anisotropic Gaussian refinement (AAGR) module. This module processes semantic features F_{12} through anatomical centroid regression and anisotropic Gaussian modeling, thereby integrating semantic attention with geometric constraints. It should be noted that no manual landmark annotation was provided for centroid supervision in our setting; the anatomical relevance of the attention focus is instead an emergent property of the classification objective, as qualitatively demonstrated in Fig. 2(e).

Anatomical Centroid Regression. Standard coordinate extraction methods (e.g., *Argmax*) are non-differentiable and suffer from quantization errors. To overcome this and enable end-to-end training, we employ the *Soft-argmax* operator [23]. This mechanism differentiably regresses the sub-pixel centroid coordinates (c_x, c_y) directly from the deepest feature map F_{12}, ensuring that the attention center aligns robustly with the semantic core of the target organ.

Anisotropic Gaussian Modeling. Instead of rigid bounding boxes, we model the region of interest as a probabilistic field using an anisotropic Gaussian distribution to accommodate irregular or elongated anatomy. The model predicts independent vertical and horizontal standard deviations (σ_w, σ_h) from the network outputs, allowing the attention window to adapt its aspect ratio to organ shape. The resulting Gaussian spatial attention map M_{gauss} models the decay of certainty from the anatomical center to the periphery:

$$M_{gauss}(x, y) = \exp\left(-\frac{(x - c_x)^2}{2\sigma_w^2} - \frac{(y - c_y)^2}{2\sigma_h^2}\right) \tag{4}$$

Finally, to confine the analysis to the valid scanning sector, we impose a physical constraint using a pre-defined Trapezoidal Mask (M_{trap}) based on the probe's geometry. The final spatial attention map is derived via the element-wise product $M_{total} = M_{gauss} \odot M_{trap}$. This operation effectively confines the feature response within the intersection of the anatomically probable region and the physically valid imaging area, filtering out irrelevant background noise with minimal parameter overhead.

2.4 Dual-Head Objective Function

To enhance the discriminative power of the learned representations, particularly for visually similar SPs, we employ a dual optimization framework with a classification head and a contrastive projection head. The classification head optimizes the standard cross-entropy loss $\mathcal{L}_{ce}$. Meanwhile, the contrastive projection head employs a supervised contrastive loss ($\mathcal{L}_{supcon}$) to increase the similarity of embeddings from the same class while pushing apart those from different classes. To further mitigate the impact of small batch sizes on contrastive learning, a first-in-first-out memory bank $\mathcal{M}$ is integrated to store historical projection embeddings [24]. By retrieving negative samples from $\mathcal{M}$, the model accesses a significantly larger distribution of negative pairs, facilitating a more compact and distinct feature space. The supervised contrastive loss is formulated as:

$$\mathcal{L}_{supcon} = -\sum_{i \in I} \frac{1}{|\mathcal{P}(i)|} \sum_{p \in \mathcal{P}(i)} \log \frac{\exp(z_i \cdot z_p/\tau)}{\sum_{\substack{k \in [A(i) \cup \mathcal{M}] \\ k \neq i}} \exp(z_i \cdot z_k/\tau)} \tag{5}$$

where $i \in I$ is the index of the anchor sample, τ is the temperature parameter, and $A(i)$ represents the current mini-batch. The terms z_i, z_p, and z_k represent the L_2-normalized projection embeddings. The set $\mathcal{P}(i)$ defines the indices of all positive samples in the combined set $\{A(i) \cup \mathcal{M}\}$ with respect to the anchor, with its cardinality $|\mathcal{P}(i)|$ serving as a normalization factor. The model is optimized by minimizing the total joint loss:

$$\mathcal{L}_{total} = \mathcal{L}_{ce} + \lambda \mathcal{L}_{supcon} \tag{6}$$

where λ is a balancing hyperparameter.

3 Experiments

3.1 Experimental Setup

Datasets and Evaluation Metrics. We utilized an abdominal ultrasound dataset consisting of nine categories, hereafter referred to as AbdoSP-9, curated by chief physicians from a hospital in accordance with routine clinical examination protocols, with written informed consent obtained from all participants. Annotation reliability was verified by independent re-labelling of 30 subjects by one chief physician and two attending physicians, with discrepancies resolved

Table 1. Distribution of samples across the AbdoSP-9 dataset

SPs	AA	CL	IVC	LPV	THV	IVC-GB	PB	PT	NSP
Number	326	380	328	544	264	391	191	369	222

by consensus, yielding Fleiss' κ=0.7114. The dataset encompasses eight clinically significant standard planes (SPs), including the Abdominal Aorta (AA), Caudate Lobe (CL), Inferior Vena Cava (IVC), Left Portal Vein (LPV), Three Hepatic Veins (THV), IVC with Gallbladder (IVC-GB), Pancreatic Body (PB), and Pancreatic Tail (PT), all of which serve as essential landmarks for liver and pancreatic assessment. Additionally, a Non-SP (NSP) category comprising transitional frames and motion blur was incorporated to enhance the model's robustness against background noise and indistinct structures frequently encountered in dynamic scanning sequences.

To assess generalization and real-world stability, we evaluated on the public FPUS23 dataset [21] for fetal growth monitoring, including Abdominal Circumference (AC), Femur Length (FL), Biparietal Diameter (BPD), and Background (No Plane), and collected six additional continuous-scanning video sequences to measure inference speed and frame-to-frame consistency in dynamic scenarios. Comprehensive comparison used accuracy, precision, recall, and F1-score, while dynamic performance on empirical scans was assessed by success rate, mean absolute error (MAE), and temporal localization bias.

Implementation Details. GCM-Net was implemented in PyTorch and trained on four NVIDIA RTX 4080 GPUs. Inputs were resized to 1024×1024 and augmented with random resizing, horizontal flipping, and color jittering. We used Adam with a learning rate of 10^{-4} and AMP. The mini-batch size was 2 with gradient accumulation of 4, and the supervised-contrastive weight was set to $\lambda = 0.6$ based on extensive experiments. The memory bank size is set to $M = 512$, and the model is trained for 10000 iterations. Inference latency on an NVIDIA RTX 4080 with FP16/TF32 mixed precision is 65.25 ms/frame, encompassing pre-processing (1024×1024 resizing and normalization), model inference, and post-processing, satisfying the real-time requirement for clinical deployment.

3.2 Quantitative Comparison

We evaluated our GCM-Net against a diverse set of benchmark methodologies, including the ultrasound-specific classification model SonoNet64 [10], leading vision foundation models such as DINOv3 [25], and cutting-edge medical-task-specific models like MedViTV2 [13].

As summarized in Table 2, GCM-Net achieved state-of-the-art performance across both datasets, reaching 99.53% accuracy on FPUS23 and 81.37% on the more complex AbdoSP-9 dataset. These results represented significant improvements, specifically outperforming DINOv3 by 3.89% on FPUS23 and surpassing

Table 2. Quantitative comparisons of different models

Method	FPUS23				AbdoSP-9			
	Acc(%)	Prec(%)	Rec(%)	F1(%)	Acc(%)	Prec(%)	Rec(%)	F1(%)
SonoNet64 [10]	94.22	94.50	94.24	94.02	58.78	53.54	54.63	53.42
DINOv3 [25]	95.64	96.00	95.75	95.75	57.25	58.67	54.75	55.56
MedViTV2-small [13]	97.91	97.92	97.91	97.90	73.44	78.79	73.44	74.31
GCM-Net(Ours)	**99.53**	**99.53**	**99.53**	**99.52**	**81.37**	**83.85**	**79.14**	**78.80**

Table 3. Ablation study on the AbdoSP-9 dataset

Setting	LMFF	AAGR	SupCon	Acc(%)	Prec(%)	Rec(%)	F1(%)
Only Backbone				72.37	69.41	68.72	66.78
w/o LMFF		✓	✓	80.61	81.94	78.87	78.27
w/o AAGR	✓		✓	78.63	81.01	73.90	72.55
w/o SupCon	✓	✓		75.88	77.34	72.45	70.19
Full Model	✓	✓	✓	**81.37**	**83.85**	**79.14**	**78.80**

the second-best MedViTV2-small by 7.93% on the AbdoSP-9 task. Furthermore, GCM-Net outperformed established benchmarks such as SonoNet-64 and DINOv3. This demonstrated superior discriminative power and robustness in capturing subtle anatomical features within complex clinical scenarios.

3.3 Ablation Study

To quantitatively evaluate the contribution of each proposed module to the model performance, we conducted a systematic ablation study on AbdoSP-9 dataset. The evaluation compared the baseline backbone with several variants produced by individually removing the LMFF module, the AAGR module, and the contrastive learning branch. Table 3 summarizes the experimental results.

Ablation results demonstrated that GCM-Net significantly outperformed the baseline, yielding improvements of 9.00% in accuracy and 12.02% in F1-score. These findings validated the synergy of the integrated modules. Contrastive learning proved to be the most critical component and its removal resulted in a 5.49% accuracy drop, indicating its essential role in enhancing feature discriminability. Furthermore, the AAGR module was vital for anatomical capture, as evidenced by a 5.24% recall decrease upon its removal. Finally, the LMFF module contributed a 1.91% precision gain through granular feature fusion.

We ablate τ and λ on AbdoSP-9. Low τ=0.1 achieves peak accuracy (81.37%) by sharpening the contrastive distribution for hard negative mining. Higher values over-smooth it (e.g., 76.03% at τ=0.5). For loss weight, λ=0.6 (81.37%) balances contrastive regularization with classification, with under-weighting (λ=0.2, 76.49%) and over-weighting (λ=1.0, 78.47%) degrading performance.

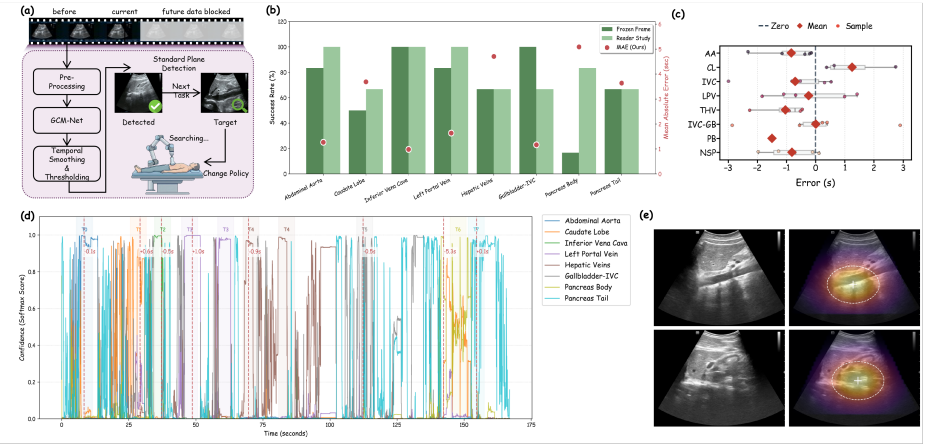

Fig. 2. Evaluation of GCM-Net for real-time causal processing of in vivo ultrasound scans. (a) Causal inference pipeline and conceptual framework for robotic ultrasound guidance. (b) Success rates and mean absolute error (MAE) of temporal localization across all anatomical planes. (c) Error distribution for successful detections shown as boxplots. (d) Temporal confidence refinement (TCR) results. (e) Representative AAGR Gaussian attention maps overlaid on ultrasound frames.

3.4 Robustness Analysis in Real-world Scenarios

Fig. 2(a) illustrates the causal inference pipeline for real-time ultrasound processing using GCM-Net. GCM-Net processes each frame at 30 fps, maintaining a W_s=30-frame moving-average confidence $\bar{c}_t^{(k)}$ per class k. After a 3.0 s warm-up, the system enters peak-tracking mode once $\bar{c}_t^{(k)}$ exceeds a validation-derived threshold θ_k, continuously updating the candidate peak t^*. Detection locks when no improvement is observed within a W_p=60-frame patience window:

$$t^* = \underset{t \geq t_{\text{warm}}}{\arg\max}\ \bar{c}_t^{(k)}, \quad s_{n,k} = \mathbf{1}[\text{TCR fires}] \cdot \mathbf{1}[|t^*_{n,k} - t^{\text{GT}}_{n,k}| \leq \delta], \tag{7}$$

where a detection is counted successful if TCR fires and the temporal error falls within δ=3.0 s.

Fig. 2(d) presents the temporal confidence refinement (TCR) result on a representative clinical scan, where temporally correlated detection peaks suppress single-frame false positives while preserving the stability needed for robust scanning guidance. Quantitatively (green bars, Fig. 2(b)), GCM-Net performs robustly across all tasks, achieving a 70.83% overall success rate and a low temporal MAE of 1.06 s among successful detections. To assess clinical utility, a reader study had a chief physician review the frames selected by GCM-Net, yielding a clinical accuracy of 81.25% (39/48 tasks, light green bars) that surpasses the mathematical metric. This discrepancy is evident in challenging tasks such as pancreatic body classification, where clinicians often rewind the video to

retrospectively select an optimal frame, causing the ground-truth timestamp (defined at the freeze-frame) to lag behind the model's real-time detection. Overall, GCM-Net's real-time causal inference shows significant potential as a high-level controller for autonomous robotic ultrasound (Fig. 2(a)) and an intelligent assistant for manual scanning.

4 Conclusion

We propose GCM-Net, an anatomy-aware framework for robust standard plane classification in abdominal ultrasound. By integrating differentiable spatial gating with foundation model adaptation and a memory-bank-based dual-head objective, GCM-Net effectively enhances feature discriminability while suppressing noise and artifacts. Evaluations on abdominal and fetal datasets demonstrate superior accuracy and robust causal video inference, establishing a scalable pathway toward standardized diagnosis and autonomous robotic ultrasound.

Acknowledgments. This work was supported by the National Key Research and Development Program of China (2025YFC2425700, 2022YFC2405200), National Natural Science Foundation of China (82472115), Beijing Natural Science Foundation (L252049, L252007) and Tsinghua University Dushi Program.

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

References

1. Leigh, N., Hammill, C.W.: Liver ultrasound: Normal anatomy and pathologic findings. Surgery Open Science 19, 178-198 (2024)
2. Zhang, T., Yan, J., Zhou, X., Wu, B., Zhang, C., Tang, M., Huang, P.: Application of Ultrasound Localization Microscopy in evaluating the type 2 diabetes progression. Cyborg and Bionic Systems 6, 0117 (2025)
3. Kamaya, A., Fetzer, D.T., Seow, J.H., et al.: LI-RADS US Surveillance Version 2024 for Surveillance of Hepatocellular Carcinoma: An Update to the American College of Radiology US LI-RADS. Radiology, 313, e240169 (2024)
4. Li, Z., Xu, Y., Zhang, L., Han, T., Yang, X., Wang, Y., Liu, M., Xin, S., Liu, L., Liao, H.: Beyond hospital reach: Autonomous lightweight ultrasound robot for liver sonography. arXiv preprint arXiv:2510.08106 (2025)
5. Liang, H., Zhang, S., Ning, G., Luo, J., Liao, H.: Quasi-static Elastography-driven Automated Robotic Ultrasound Screening and Localization. IEEE Transactions on Biomedical Engineering 1-14 (2025)
6. Han, T., Ning, G., Liang, H., Li, Z., Jiang, Z., Chen, F., Kang, Y., Luo, J., Liao, H.: Toward Clinical Applications of Intelligent Robotic Ultrasound Systems. IEEE Reviews in Biomedical Engineering, 19, 227–247 (2026)
7. Wu, Z., Wang, X., Cao, Y., Zhang, W., Xu, Q.: Robotic Ultrasound Scanning End-Effector with Adjustable Constant Contact Force. Cyborg and Bionic Systems 6, 0251 (2025)

8. Chen, H., Dou, Q., Ni, D., Cheng, J.-Z., Qin, J., Li, S., Heng, P.-A.: Automatic Fetal Ultrasound Standard Plane Detection Using Knowledge Transferred Recurrent Neural Networks. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, pp. 507-514. Springer International Publishing, (2015)
9. Maraci, M.A., Napolitano, R., Papageorghiou, A., Noble, J.A.: Searching for Structures of Interest in an Ultrasound Video Sequence. In: Machine Learning in Medical Imaging, pp. 133-140. Springer International Publishing, (2014)
10. Baumgartner, C.F., Kamnitsas, K., Matthew, J., Fletcher, T.P., Smith, S., Koch, L.M., Kainz, B., Rueckert, D.: SonoNet: Real-Time Detection and Localisation of Fetal Standard Scan Planes in Freehand Ultrasound. IEEE Trans Med Imaging 36, 2204–2215 (2017)
11. Lawley, A., Hampson, R., Worrall, K., Dobie, G.: Analysis of neural networks for routine classification of sixteen ultrasound upper abdominal cross sections. Abdominal Radiology 49, 651-661 (2024)
12. Yu, T., Tsui, P.H., Leonov, D., Wu, S., Bin, G., Zhou, Z.: LPC-SonoNet: A Lightweight Network Based on SonoNet and Light Pyramid Convolution for Fetal Ultrasound Standard Plane Detection. Sensors (Basel) 24, (2024)
13. Nejati Manzari, O., Asgariandehkordi, H., Koleilat, T., Xiao, Y., Rivaz, H.: Medical image classification with KAN-integrated transformers and dilated neighborhood attention. Applied Soft Computing 186, 114045 (2026)
14. Tehsin, S., Alshaya, H., Bouchelligua, W., Nasir, I.M.: Hybrid State-Space and Vision Transformer Framework for Fetal Ultrasound Plane Classification in Prenatal Diagnostics. Diagnostics (Basel) 15, (2025)
15. Chowdary, G.J., Yin, Z.: Med-former: A transformer based architecture for medical image classification. In: International conference on medical image computing and computer-assisted intervention, pp. 448-457. Springer, (2024)
16. Fiorentino, M.C., Villani, F.P., Di Cosmo, M., Frontoni, E., Moccia, S.: A review on deep-learning algorithms for fetal ultrasound-image analysis. Medical Image Analysis 83, 102629 (2023)
17. Cohen, H.L., McGahan, J.P., Hertzberg, B.S., Meilstrup, J.W., Needleman, L., Hashimoto, B.E., Foley, W.D., Townsend, R.R., Frates, M., Bromley, B.: AIUM practice guideline for the performance of an ultrasound examination of the abdomen and/or retroperitoneum. Journal of Ultrasound in Medicine 27, 319-326 (2008)
18. Conti, E., Rosati, R., Federici, L., Mancini, A., Fiorentin, M.C.: Challenging DINOv3 Foundation Model under Low Inter-Class Variability: A Case Study on Fetal Brain Ultrasound. arXiv preprint arXiv:2511.01915 (2025)
19. Meyer, A., Murali, A., Zarin, F., Mutter, D., Padoy, N.: Ultrasam: a foundation model for ultrasound using large open-access segmentation datasets. International Journal of Computer Assisted Radiology and Surgery (2025)
20. Jiao, J., Zhou, J., Li, X., Xia, M., Huang, Y., Huang, L., Wang, N., Zhang, X., Zhou, S., Wang, Y., Guo, Y.: USFM: A universal ultrasound foundation model generalized to tasks and organs towards label efficient image analysis. Medical Image Analysis 96, 103202 (2024)
21. Prabakaran, B.S., Hamelmann, P., Ostrowski, E., Shafique, M.: FPUS23: An Ultrasound Fetus Phantom Dataset With Deep Neural Network Evaluations for Fetus Orientations, Fetal Planes, and Anatomical Features. IEEE Access 11, 58308-58317 (2023)
22. Wu, J., Wang, Z., Hong, M., Ji, W., Fu, H., Xu, Y., Xu, M., Jin, Y.: Medical SAM adapter: Adapting segment anything model for medical image segmentation. Medical Image Analysis 102, 103547 (2025)

23. Nibali, A., He, Z., Morgan, S., Prendergast, L.: Numerical coordinate regression with convolutional neural networks. arXiv preprint arXiv:1801.07372 (2018)
24. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised Feature Learning via Non-parametric Instance Discrimination. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3733-3742. (2018)
25. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)