SEMC: Structure-Enhanced Mixture-of-Experts Contrastive Learning for Ultrasound Standard Plane Recognition

Anonymous submission

Abstract

Ultrasound standard plane recognition is essential for clinical tasks such as disease screening, organ evaluation, and biometric measurement. However, existing methods fail to effectively exploit shallow structural information and struggle to capture fine-grained semantic differences through contrastive samples generated by image augmentations, leading to poor recognition of structural and discriminative details in ultrasound standard planes. To address these issues, we propose Structure-Enhanced Mixture-of-Experts Contrastive Learning (SEMC), a novel framework that combines structure-aware feature fusion with expert-guided contrastive learning. Specifically, we propose a Semantic-Structure Fusion Module (SSFM) to exploit multi-scale structural information and enhance the model's ability to perceive fine-grained structural details by effectively aligning shallow and deep features. Meanwhile, a Mixture-of-Experts Contrastive Recognition Module (MCRM) is designed to perform hierarchical contrastive learning and classification across multi-level features using a mixture-of-experts (MoE) mechanism, further improving class separability and overall recognition performance. More importantly, we also curate a large-scale and meticulously annotated liver ultrasound dataset containing six standard planes. Extensive experimental results on our in-house dataset and two public datasets demonstrate that SEMC outperforms recent state-of-the-art methods across various metrics.

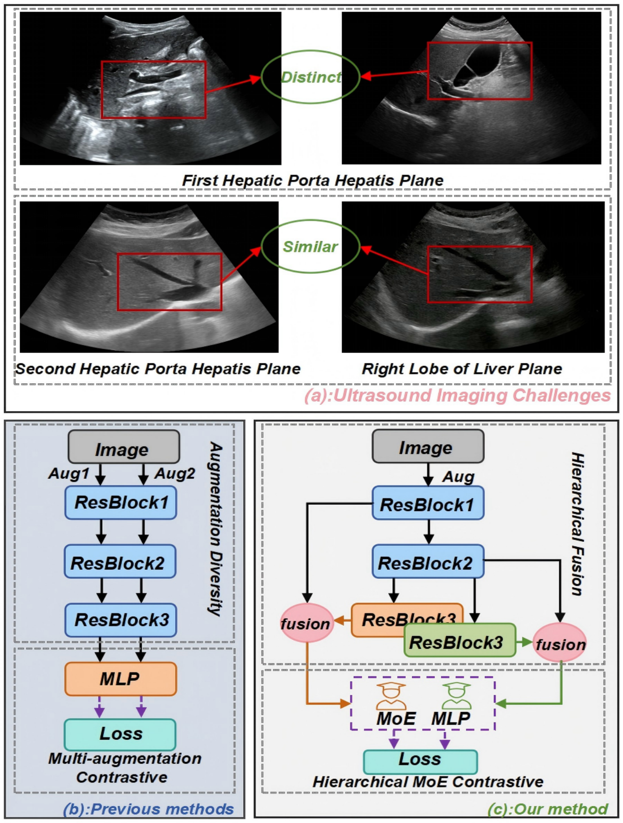

Figure 1: (a) Illustrates the challenges in ultrasound images, featuring large intra-class variation and high inter-class similarity. (b) Previous methods mainly rely on deep semantic features, neglecting shallow structural cues. (c) Our SEMC framework integrates shallow structure via semantic-structure fusion and employs a MoE for hierarchical contrastive learning, yielding more discriminative and structure-aware representations.

Introduction

Ultrasound imaging is one of the most widely used medical imaging techniques in clinical practice due to its non-invasive nature, real-time capability, high efficiency, and low cost (Spencer and Adler 2008). It is particularly effective for visualizing human organs, soft tissues, and for conducting prenatal examinations. In clinical settings, acquiring standard planes is essential for accurate diagnosis and quantitative analysis. These planes provide clinicians reliable structural visualization and consistent reference points for measurement (Wang et al. 2022; Di Cosmo et al. 2022). For instance, in prenatal ultrasound, the femoral standard plane, thalami standard plane, and abdominal standard plane are commonly used to measure fetal length, head circumference, and abdominal circumference, respectively. These measurements are important indicators for fetal growth assessment (Salomon et al. 2006; Guo et al. 2022). In practice, the quality of selected standard planes varies depending on the operator's experience and scanning technique, which can impact the accuracy of growth measurements and clinical decisions (Salomon et al. 2011; Maraci et al. 2014).

Recent studies have explored deep learning-based methods for standard plane (SP) identification and have achieved promising results (Pu et al. 2021; Migliorelli et al. 2024). However, SP recognition still encounters several critical challenges. Ultrasound imaging quality varies significantly due to speckle noise, low contrast, and indistinct boundaries, which makes structural region detection particularly difficult. As shown in Figure 1(a), inter-class differences between SP are often subtle (Lin et al. 2019; Xie et al. 2020). For example, certain liver SP differ only slightly in shape or texture, requiring fine-grained discrimination (Baumgartner et al. 2016, 2017). Additionally, large intra-class variations caused by differences in acquisition angles, probe pressure, and operator experience can lead to inconsistent recognition performance (Yu et al. 2024; Krishna and Kokil 2024). Most existing approaches emphasize deep semantic representations while neglecting the extraction of shallow structural information, thereby limiting the model's ability in both semantic discrimination and structural perception (Men et al. 2023; Li et al. 2025). Furthermore, although contrastive learning has been introduced as an auxiliary strategy by constructing augmented positive and negative pairs, such methods often struggle to effectively capture the fine-grained semantic distinctions inherent in ultrasound images, as shown in Figure 1(b).

To address these issues, we propose a novel framework, **Structure-Enhanced Mixture-of-Experts Contrastive Learning (SEMC)**, which effectively integrates structure-aware feature fusion with expert-guided contrastive learning to tackle the challenges inherent in ultrasound standard plane recognition. In this framework, we design a dedicated Semantic-Structure Fusion Module (SSFM) that explicitly aligns and integrates shallow structural cues with deep semantic representations. This enhances the model's sensitivity to fine-grained structural details. To further enhance the model's discriminative capability, SEMC designs a Mixture-of-Experts Contrastive Recognition Module (MCRM), in which multiple expert branches are specifically designed to specialize in different aspects of the feature space and collaboratively perform hierarchical contrastive learning, as illustrated in Figure 1(c). By enforcing contrastive objectives at multiple feature levels, the framework promotes improved inter-class separability and more compact intra-class clustering within the representation space. In addition, we construct a high-quality liver ultrasound dataset named **LP2025** containing six standard planes, aiming to address the lack of publicly available data in the field of standard plane recognition. Evaluations on this dataset and two public standard plane benchmark datasets demonstrate that SEMC outperforms existing state-of-the-art methods across multiple metrics, showing strong potential for clinical application. In summary, our main contributions are as follows:

- We propose Structure-Enhanced Mixture-of-Experts Contrastive Learning (SEMC), which integrates a Semantic-Structure Fusion Module (SSFM) and a Mixture-of-Experts Contrastive Recognition Module (MCRM) to enhance fine-grained structural perception and discriminative feature representation for ultrasound plane recognition.
- We construct LP2025, a liver ultrasound dataset comprising six standard planes, aiming to address the shortage of liver-specific standard plane benchmarks in the field.
- Extensive experimental results on two public datasets and an in-house liver ultrasound dataset demonstrate that our SEMC framework consistently outperforms existing methods in standard plane recognition tasks.

Related Work

Standard Plane Recognition in Ultrasound

Standard plane recognition is a fundamental task in medical image understanding, with broad clinical applications such as disease screening, organ function assessment, and biometric measurement. Early methods relied on handcrafted features combined with traditional classifiers (e.g., SVM, KNN), but their generalization capability was limited due to weak feature representation and poor image quality (Christodoulou et al. 2003; Latha, Samiappan, and Kumar 2020; Huang et al. 2020; Liao, Cheng, and Chan 2024). In recent years, convolutional neural networks (CNN) have emerged as the mainstream solution. Notably, the SonoNet (Baumgartner et al. 2017) series, based on the VGG architecture, achieved promising results in automatic recognition of fetal standard planes. Subsequent works have incorporated strategies such as multi-task learning, attention mechanisms, and structural priors to enhance the model's ability to perceive key regions and fine-grained variations (Cai et al. 2018; Zhu, Salcudean, and Rohling 2022; Yu et al. 2024; Ciobanu et al. 2025). However, these methods primarily focus on high-level semantic modeling, often overlooking the importance of shallow structural information and spatial context in the decision-making process. As a result, their performance tends to degrade in real-world scenarios characterized by subtle inter-class differences and complex backgrounds.

Contrastive Learning and Mixture-of-Experts

Contrastive learning has shown great potential in improving representation learning, especially in medical imaging tasks with limited data and ambiguous class boundaries. Methods like MoCo (Sowe 2025) and SimCLR (Chen et al. 2020) optimize the feature space by constructing positive and negative sample pairs, promoting intra-class compactness and inter-class separability. Supervised contrastive learning (Khosla et al. 2021; Lin et al. 2024) further improves discriminability and semantic consistency by using label information during pair construction. The MoE paradigm has gained attention for its dynamic modeling capabilities and parameter efficiency (Shazeer et al. 2017; Riquelme et al. 2021; Zoph 2022). Models such as Conditional MoE (Zhu et al. 2022) and Switch Transformer (Fedus, Zoph, and Shazeer 2022) have achieved breakthroughs in natural language processing and computer vision. However, in ultrasound image analysis, where structures are complex and boundaries often blurred, the integration of MoE and contrastive learning is still underexplored. Specifically, effective mechanisms to guide expert collaboration through structural cues are lacking. To address these challenges, we propose a novel framework that integrates structure-enhanced feature fusion with contrastive expert modeling. Our method explicitly enhances the model's ability to recognize fine-grained differences in standard plane recognition tasks.

LP2025 Dataset Construction

Data Collection

To advance research on standard plane recognition in liver ultrasound imaging, we present a comprehensive and high-quality dataset, **LP2025**, specifically curated for deep learning-based anatomical understanding and classification tasks. This dataset provides a reliable benchmark for evaluating model performance on both fine-grained structural recognition and real-world generalization under diverse clinical conditions. The LP2025 dataset was constructed under the clinical supervision of experienced radiologists and certified sonographers at a leading tertiary medical center. All ultrasound examinations were conducted using high-resolution diagnostic ultrasound systems to ensure optimal image quality and structural clarity. A standardized imaging protocol was employed to unify scanning procedures across different patients and sessions, ensuring consistency in anatomical coverage, image resolution, and diagnostic relevance.

Each subject underwent a systematic abdominal ultrasound scan, during which six clinically meaningful liver standard planes were meticulously captured. These planes were selected based on their diagnostic importance in hepatobiliary evaluations and their frequent usage in routine clinical workflows. The six standard planes in LP2025 are as follows:

- First Porta Hepatis Plane (FHP1): Captures the bifurcation of the portal vein, serving as a key landmark for hepatic segmentation.
- Second Porta Hepatis Plane (FHP2): Displays the continuation of the portal vein and hepatic artery, facilitating vascular assessments.
- Left Lobe Plane (LLP): Highlights the morphology and parenchymal pattern of the left hepatic lobe.
- Right Lobe Plane (RLP): Visualizes the texture and size of the right hepatic lobe, often used to assess hepatomegaly and hepatic lesions.
- Sagittal Plane of Left Portal Vein (LPV-S): Offers a clear sagittal view of the left portal vein branch, aiding in vascular diagnosis.
- Hepatorenal Plane (HRP): Shows the interface between the liver and right kidney, commonly used to detect ascites or space-occupying lesions.

In addition to the six standard planes, LP2025 includes a Non-Standard Plane (NSP) category, consisting of images that do not conform to the above-defined diagnostic views but are frequently observed in routine ultrasound examinations. The inclusion of this category introduces realistic variation and classification ambiguity, thereby enhancing model robustness and better simulating real-world clinical deployment scenarios.

Dataset Composition and Annotation Quality

To ensure the accuracy, consistency, and clinical validity of the labels, each image in the LP2025 dataset was independently annotated by a team of senior sonographers, all of whom possess over five years of hands-on experience in liver ultrasound imaging. The annotation process focused on two key aspects: standard plane classification and the presence of clearly identifiable anatomical structures.

A rigorous multi-stage quality control pipeline was implemented to maintain high annotation standards:

- Initial Review: Each annotation was independently cross-checked by a second sonographer to catch any inconsistencies or errors.
- Consensus Verification: For cases with annotation discrepancies, a discussion among at least three senior sonographers was conducted to reach a consensus, ensuring clinical agreement.
- Final Validation: A final round of inspection was performed to assess the clinical relevance of each image and to exclude low-quality or ambiguous samples that may negatively impact model training or evaluation.

This comprehensive and multi-expert review process ensures the reliability and trustworthiness of the dataset, making it a strong foundation for both algorithm development and clinical application research.

Plane	**Abbreviation**	**Number**
First Porta Hepatis Plane	FHP1	979
Second Porta Hepatis Plane	FHP2	324
Left Lobe Plane	LLP	1038
Right Lobe Plane	RLP	490
Sagittal Plane of Left Portal Vein	LPV-S	840
Hepatorenal Plane	HRP	1072
Non-Standard Plane	NSP	4626
Total	–	9369

Table 1: Distribution of images in the LP2025 dataset by standard plane

The final curated version of the LP2025 dataset consists of 9,369 high-quality, well-annotated liver ultrasound images. These images are distributed across six clinically recognized standard planes. The detailed distribution is presented in Table 1. The LP2025 dataset will be made publicly available for non-commercial academic research upon request.

Methodology

Overview

We propose a Structure-Enhanced Mixture-of-Experts Contrastive Learning (SEMC) framework for ultrasound standard plane recognition. As illustrated in Figure 2, our core contributions include two main components: (1) A semantic-structure fusion module, which explicitly aligns and integrates multi-level features to enhance structural awareness; (2) An MoE-based contrastive recognition module, which

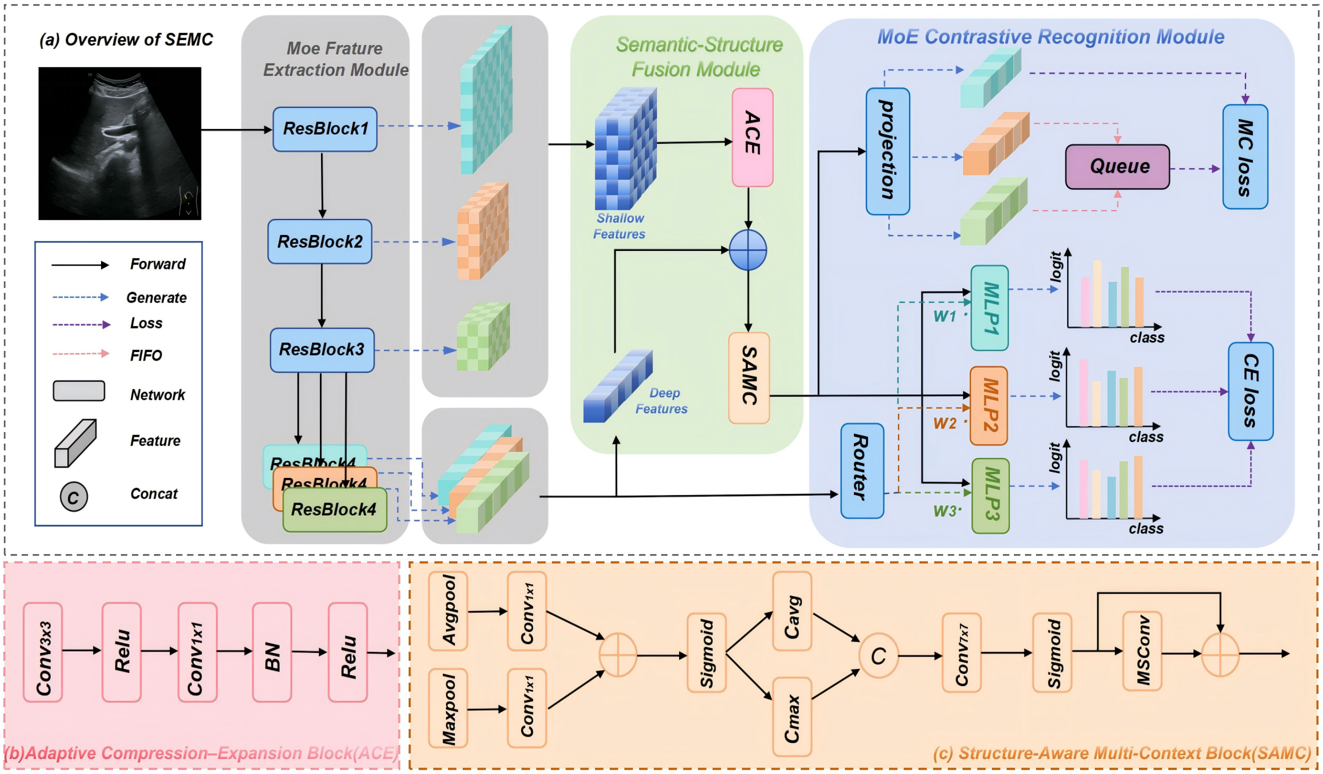

Figure 2: Architecture of the proposed SEMC framework. The architecture first employs an MoE-based feature extractor to obtain multi-level expert features from the input ultrasound image. These features are then aligned and enhanced through a Semantic-Structure Fusion Module (SSFM). The resulting representations are fed into the Mixture-of-Experts Contrastive Recognition Module (MCRM), which consists of two task branches: a multi-class classification head based on MoE as the primary task and a contrastive learning branch based on MoE as an auxiliary task, which further improves the performance of the primary task by optimizing the learned feature representations.

leverages expert-specific features for both contrastive learning and classification. By jointly optimizing the feature space of supervised and self-supervised learning, this module significantly improves the class separability and recognition performance of the model.

Mixture-of-Experts Feature Extraction Module

To effectively model fine-grained variations and complex anatomical structures in ultrasound images, we design an MoE feature extraction module based on a modified ResNet backbone. Specifically, the first three blocks (i.e., `layer1` to `layer3`) are shared across all branches and serve as a common encoder to extract low- and mid-level features. After `layer3`, we introduce three parallel, structurally identical but parameter-independent fourth blocks (denoted as `layer4-1`, `layer4-2`, and `layer4-3`), forming three specialized deep expert paths:

$$F_1, F_2, F_3 = \mathrm{ResNet}_{1\sim 3}(x), \tag{1}$$

$$D_1 = \mathrm{ResNet}_{4-1}(F_3), \tag{2}$$

$$D_2 = \mathrm{ResNet}_{4-2}(F_3), \tag{3}$$

$$D_3 = \mathrm{ResNet}_{4-3}(F_3), \tag{4}$$

where F_3 is the output feature of the shared backbone, and $\{D_1, D_2, D_3\}$ represents the high-level semantic features extracted by each expert path. This design introduces diverse feature representations through decoupled expert parameters, which facilitates the modeling and selection of different semantic perspectives in subsequent fusion modules.

Semantic-Structure Fusion Module (SSFM)

Most existing methods focus only on deep features while neglecting the complementary value of shallow features, especially when anatomical contrast is weak or boundaries are unclear. To address this limitation, we propose the Structure-Semantic Fusion Module (SSFM), which integrates shallow and deep features through two components: the Adaptive Compression–Expansion (ACE) Block and the Structure-Aware Multi-Context (SAMC) Block. This design enhances both feature discrimination and structural representation.

Adaptive Compression–Expansion Block (ACE): To address spatial and channel mismatches between shallow and deep features, we propose a lightweight ACE module. ACE aligns shallow features $\{F_1, F_2, F_3\}$ with deep expert features through progressive downsampling and channel adaptation. Each ACE block processes input $\mathbf{X}_0 = F_i$

through L stages:

$$\mathbf{X}_{i+1} = \mathrm{BN}\Big(\mathrm{Conv}_{1\times 1}\big(\mathrm{ReLU}\big(\mathrm{BN}\cdot \mathrm{DWConv}^{s=2}_{3\times 3}(\mathbf{X}_i)\big)\big)\Big), \quad i = 0, 1, \ldots, L-1, \tag{5}$$

where channel size doubles at each step: $C_i = C_{\text{in}} \times 2^i$. After L stages, a final 1×1 convolution with BN and ReLU projects $\mathbf{X}_L$ to the target channel count C_{out}:

$$F_i' = \mathrm{ReLU}\left(\mathrm{BN}\left(\mathrm{Conv}_{1\times 1}(\mathbf{X}_L)\right)\right), \tag{6}$$

yielding aligned features $F_i' \in \mathbb{R}^{C_{\text{out}} \times H_L \times W_L}$.

ACE first reduces spatial resolution via strided depthwise convolutions, then adjusts channels with pointwise convolutions, ensuring efficiency and structural preservation. Aligned features are fused with deep expert features $\{D_1, D_2, D_3\}$ via element-wise addition:

$$M_i = F_i' + D_i, \quad i = 1, 2, 3. \tag{7}$$

Compared to concatenation, this strategy avoids channel redundancy, reduces parameters, and promotes learning of shared discriminative patterns.

Structure-Aware Multi-Context Block (SAMC): To enhance the discriminative power and structural representation of fused features, we propose the SAMC. This module uses channel expansion and multi-scale convolutions to reconstruct features, applies channel-spatial attention to highlight important regions, and employs residual connections to retain information. For each fusion branch $\mathcal{M}_i$, the processing proceeds as follows:

$$\mathbf{C}_i = \sigma\Big(\mathrm{FC}_2\big(\delta(\mathrm{FC}_1(\mathrm{AvgPool}(\mathcal{M}_i)))\big) + \mathrm{FC}_2\big(\delta(\mathrm{FC}_1(\mathrm{MaxPool}(\mathcal{M}_i)))\big)\Big), \quad i = 1, 2, 3, \tag{8}$$

where $\mathbf{C}_i \in \mathbb{R}^C$ denotes the adaptive channel attention. Global pooling captures context, while shared FC layers with activation $\delta(\cdot)$ model dependencies. The sigmoid $\sigma(\cdot)$ outputs normalized weights.

$$\mathbf{S}_i = \sigma\left(\mathrm{Conv}\left([\mathrm{Mean}(\mathbf{C}_i \odot \mathcal{M}_i, \dim = 1), \mathrm{Max}(\mathbf{C}_i \odot \mathcal{M}_i, \dim = 1)]\right)\right), \quad i = 1, 2, 3, \tag{9}$$

where $\mathbf{S}_i \in \mathbb{R}^{1\times H\times W}$ is the spatial attention map derived from pooled channel-attended features. Convolution integrates spatial cues, guiding the network to focus on key regions.

$$\mathbf{O}_i = \mathrm{PConv}_2\Big(\mathrm{Shuffle}\big(\mathrm{Concat}(\{\mathbf{F}_k^{(i)}\}_{k=1}^K)\big)\Big), \quad i = 1, 2, 3, \tag{10}$$

where $\{\mathbf{F}_k^{(i)}\}_{k=1}^K$ are multi-scale features from the spatially enhanced input $\mathbf{S}_i \odot \mathbf{C}_i \odot \mathcal{M}_i$. These are concatenated, shuffled for channel interaction, and compressed via pointwise convolution to generate the fused output $\mathbf{O}_i$.

MoE Contrastive Recognition Module (MCRM)

Existing methods have introduced contrastive learning as an auxiliary recognition strategy by constructing augmented positive and negative sample pairs for training. However, such approaches often struggle to effectively capture the inherent fine-grained semantic variations in ultrasound images. To address this, SEMC designs an MCRM composed of two synergistic branches. (1) MoE Contrastive Branch: Multiple expert sub-networks focus on different aspects of the feature space and collaboratively perform hierarchical contrastive learning. This design enhances inter-class separability and intra-class compactness in the learned representations. (2) MoE Recognition Branch: Different expert subnetworks extract discriminative information from diverse perspectives and scales, thereby improving the model's ability to accurately recognize various standard planes.

MoE Contrastive Branch: To fully utilize multi-level semantic and spatial information for ultrasound plane recognition, we propose a MoE Enhanced Contrastive Branch with three expert branches sharing a backbone but supervised by different fusion views. The first expert output $\mathbf{O}_1$ serves as the contrastive anchor (query), while $\mathbf{O}_2$ and $\mathbf{O}_3$ act as positive keys to update a dynamic queue for negative sampling. We concatenate classification logits from all experts and replicate labels for semantic supervision. The current expert features are combined with a momentum memory queue $\mathcal{Q}$ storing historical features for structural contrastive learning:

$$\mathbf{O}_{\text{con}} = \mathrm{Concat}(\mathbf{O}_1, \mathbf{O}_2, \mathbf{O}_3, \mathcal{Q}), \tag{11}$$

$$\mathbf{Y}_{\text{con}} = \mathrm{Concat}(\mathbf{y}, \mathbf{y}, \mathbf{y}, \mathbf{y}_{\mathcal{Q}}), \tag{12}$$

where $\mathbf{O}_i$ are expert outputs, $\mathbf{y}$ and $\mathbf{y}_{\mathcal{Q}}$ are labels for the current batch and queue. The concatenation along the batch dimension enables more effective contrastive learning.

We define two complementary losses: a supervised contrastive loss that uses label information to bring semantically similar samples closer, and a self-supervised contrastive loss that identifies positive pairs by mining class-consistent samples from the batch and queue without relying on labels:

$$\mathcal{L}_{\text{sup}} = \mathrm{SupCon}(\mathbf{O}_{\text{con}}, \mathbf{Y}_{\text{con}}), \tag{13}$$

$$\mathcal{L}_{\text{self}} = \mathrm{SelfCon}(\mathbf{O}_{\text{con}}). \tag{14}$$

The final objective integrates both losses with a balancing factor λ:

$$\mathcal{L}_{\text{mc}} = \mathcal{L}_{\text{sup}} + \lambda \mathcal{L}_{\text{self}}, \tag{15}$$

where λ controls the relative strength of the self-supervised signal. This unified contrastive framework jointly optimizes explicit semantic discrimination and implicit structural alignment, enhancing representation robustness and generalization in ultrasound standard plane recognition.

MoE Recognition Branch: In standard plane recognition, expert subnetworks capture diverse features, but simple averaging may ignore sample-specific differences. To address this, we introduce a MoE classification branch with a learnable sparse gating mechanism. Using Gumbel-Softmax (Lin et al. 2017), our gate adaptively selects informative experts while remaining differentiable, improving robustness to ambiguity and variation.

	FPUS23				CAMUS			
Method	Accuracy↑	Precision↑	Recall↑	F1↑	Accuracy↑	Precision↑	Recall↑	F1↑
Diffmic (Yang et al. 2023)	95.29	80.63	81.58	81.08	80.91	80.25	79.20	79.69
Area (Chen et al. 2023)	95.20	93.37	95.71	94.40	81.59	80.17	**82.53**	80.88
Shike (Jin et al. 2023)	95.15	93.80	94.93	94.31	80.48	79.58	79.79	79.52
Metaformer (Yu et al. 2023)	95.52	94.19	94.48	94.53	81.52	81.58	79.69	80.49
Cast (Ke, Mo, and Yu 2024)	95.24	94.23	93.99	94.43	81.34	80.85	79.88	80.31
Supmin (Mildenberger et al. 2025)	95.28	94.03	94.68	94.34	81.13	80.88	78.84	79.71
Ours	**95.78**	**94.38**	**95.81**	**95.06**	**82.13**	**82.03**	80.08	**80.93**

Table 2: Quantitative comparison of different models on the FPUS23 and CAMUS datasets.

Let the semantic-structural feature be $\mathbf{O} \in \mathbb{R}^{B\times C\times H\times W}$, which is input to a lightweight gating network to generate expert logits $\mathbf{l} \in \mathbb{R}^{B\times N}$. This network applies adaptive average pooling followed by a linear layer. Gating weights $\mathbf{w}$ are obtained via the Gumbel-Softmax function:

$$\mathbf{w} = \text{GumbelSoftmax}(\mathbf{l}, \tau), \tag{16}$$

where τ controls distribution sparsity. Given expert predictions $\mathbf{z}_n \in \mathbb{R}^{B\times C}$, we stack them as $\mathbf{Z} \in \mathbb{R}^{B\times N\times C}$ and compute the fused output as:

$$\mathbf{z}_{\text{fused}} = \sum_{n=1}^{N} w_n \cdot \mathbf{z}_n. \tag{17}$$

The fused prediction $\mathbf{z}_{\text{fused}}$ is supervised using a standard cross-entropy loss:

$$\mathcal{L}_{\text{moe}} = \text{CE}(\mathbf{z}_{\text{fused}}, \mathbf{y}). \tag{18}$$

This design improves semantic classification flexibility and reduces redundant expert collaboration, enhancing generalization. To balance the main classification loss L_{moe} and the contrastive loss L_{mc}, we use a lightweight adaptive network. Given input feature $\mathbf{O} \in \mathbb{R}^{B\times C\times H\times W}$, it predicts a sample-specific weight $\alpha = g(\mathbf{O}) \in (0, 1)$, and the total loss is computed as a weighted combination of the two:

$$L_{\text{total}} = \alpha \cdot L_{\text{moe}} + (1 - \alpha) \cdot L_{\text{mc}}. \tag{19}$$

The balancing factor α is adaptively adjusted according to sample difficulty or training stage dynamics. This mechanism eliminates the need for manually set hyperparameters, enabling end-to-end learning of optimal weights. Consequently, it improves the stability and effectiveness of multi-task collaborative training.

Experimental Results

Datasets & Metrics

FPUS23 (Prabakaran et al. 2023) is a public fetal ultrasound dataset for standard plane recognition, covering key views such as head, abdomen, femur, and thorax, with expert annotations suitable for supervised learning. **CAMUS** (Leclerc et al. 2019) is a cardiac ultrasound dataset originally for segmentation; we selected apical two- and four-chamber views and annotated them for classification. Its subject diversity supports generalization evaluation. Additionally, we use our in-house **LP2025** dataset, containing standard liver planes for abdominal ultrasound analysis.

To fairly compare our method with others, we adopt four commonly used classification metrics: Accuracy, Precision, Recall, and F1-score.

Implementation Details

Our proposed method is implemented based on the PyTorch framework. All experiments are conducted on a server equipped with an NVIDIA RTX 3090 GPU, using the Python 3.8 environment. For data preprocessing, all input images from the datasets are resized to 512×512, and multiple data augmentation techniques are employed, including random rotation, horizontal and vertical flipping, and brightness adjustment. The model is optimized using stochastic gradient descent (SGD) with momentum 0.9 and weight decay 1×10^{-4}. The initial learning rate is set to 1×10^{-3} and decayed following a cosine annealing schedule. The batch size is 16, and training is conducted for up to 200 epochs.

Comparison with State-of-the-Art

To evaluate the classification performance of our model on the ultrasound standard plane recognition task, we conducted a comparative analysis with several state-of-the-art methods. These include Diffmic (Yang et al. 2023), which has demonstrated strong performance in medical image analysis; Area (Chen et al. 2023) and Shike (Jin et al. 2023), which are CNN-based and show notable improvements; MetaFormer (Yu et al. 2023) and Cast (Ke, Mo, and Yu 2024), which are Transformer-based and achieve promising results; and SupMin (Mildenberger et al. 2025), which incorporates improved supervised contrastive learning. For fair comparison, all models were trained with similar settings and evaluated on the same dataset under consistent conditions.

Results on FPUS23 Dataset: Table 2 shows that on the FPUS23 fetal standard plane dataset, our method achieves the highest performance in all metrics. We reach an Accuracy of 95.78%, outperforming the second-best MetaFormer (95.52%) by 0.26%. Our F1-score is also the highest at 95.06%, exceeding Area (94.40%) and SupMin (94.34%). These improvements come from our Semantic-Structure Fusion Module, which captures both shallow structural and

	LP2025			
Method	Accuracy↑	Precision↑	Recall↑	F1↑
Diffmic	80.07	75.35	77.81	76.27
Area	80.39	75.43	76.21	75.04
Shike	80.26	75.76	77.63	76.19
Metaformer	80.13	77.10	77.00	76.44
Cast	80.86	74.87	79.59	77.00
Supmin	80.92	75.95	78.12	76.77
Ours	**82.30**	**78.11**	**80.92**	**79.32**

Table 3: Quantitative comparison of different models on in-house LP2025 dataset.

deep semantic features effectively. Additionally, the cross-expert collaborative classification branch adaptively fuses multiple expert predictions, improving robustness and classification accuracy.

Results on CAMUS Dataset: Table 2 shows that our SEMC framework consistently outperforms existing methods on the CAMUS cardiac standard plane dataset. SEMC achieves the highest Accuracy of 82.13%, surpassing MetaFormer (81.52%) and Area (81.59%). It also obtains the best F1-score of 80.93%, outperforming CAST (80.31%) and SupMin (79.71%). These results demonstrate SEMC's effectiveness in handling large intra-class variability and high inter-class similarity common in ultrasound imaging.

Results on LP2025 Dataset: Table 3 shows that on our LP2025 liver standard plane dataset, our method achieves state-of-the-art results across all metrics. It attains an Accuracy of 82.30%, outperforming Diffmic (80.07%) by 2.23%, and an F1 score of 79.32%, exceeding SupMin (76.77%) by 2.55%. These improvements highlight our model's strong generalization in capturing discriminative structural differences and key semantic regions across diverse liver views, leading to better classification performance.

ACE	SAMC	L_{mc}	Accuracy ↑	F1 ↑
✗	✗	✗	80.26	76.98
✓	✗	✗	81.38	77.82
✓	✓	✗	81.51	77.91
✓	✗	✓	81.78	78.65
✓	✓	✓	**82.30**	**79.32**

Table 4: Ablation study of ACE, SAMC, and L_{mc} on the LP2025 dataset.

Ablation Study

Ablation Study of Each Component. We conducted ablation studies on the LP2025 dataset to assess the contribution of each module. As shown in Table 4, we progressively added the SSFM and MCRM components to the baseline. The SSFM includes the ACE and SAMC submodules, while MCRM introduces a contrastive loss (L_{mc}) to guide expert branches toward complementary representations. The baseline achieves 80.26% accuracy and 76.98% F1-score. Adding ACE improves accuracy to 81.38% and F1 to 77.82%, validating its effectiveness in capturing shallow structure. Adding SAMC on top of ACE yields 81.51% accuracy and 77.91% F1, further enhancing attention to spatial regions. Alternatively, ACE combined with L_{mc} boosts accuracy to 81.78% and F1 to 78.65%, demonstrating the benefit of expert-guided contrastive learning. With all modules enabled, the model achieves the best performance: 82.30% accuracy and 79.32% F1. These results confirm that the three components are complementary and jointly improve both representation and generalization.

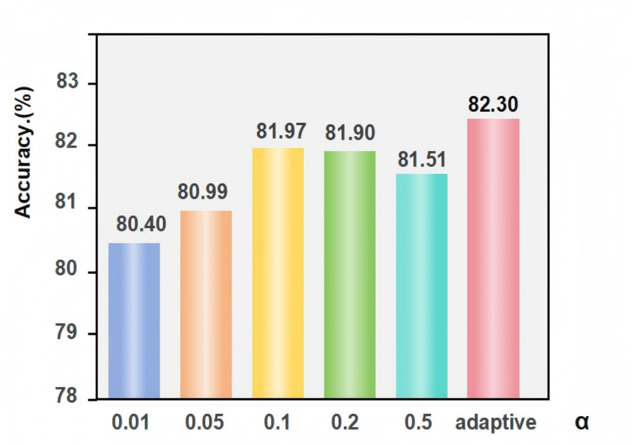

Figure 3: Performance comparison for different values of the hyperparameter α on the LP2025 dataset.

Ablation Study of Adaptive Parameter: Figure 3 presents a sensitivity analysis of the hyperparameter α defined in Equation (19). We conduct systematic ablation experiments on the LP2025 dataset, testing fixed values α of 0.01, 0.05, 0.1, 0.2, and 0.5, alongside our proposed adaptive coefficient. The results demonstrate that the adaptive strategy outperforms all fixed settings, significantly improving classification accuracy while greatly improving training stability and overall model robustness.

Conclusion

We propose a novel Structure-Enhanced Mixture-of-Experts Contrastive Learning framework (SEMC) for ultrasound standard plane recognition. SEMC incorporates a Semantic-Structure Fusion Module (SSFM) to align and integrate cross-layer features, and introduces a Mixture-of-Experts Contrastive Recognition Module (MCRM), where multiple expert branches are specifically designed to specialize in different aspects of the feature space and collaboratively perform hierarchical contrastive learning. This design effectively captures fine-grained structural details and enhances discriminative representations. In addition, we construct and carefully annotate a liver ultrasound dataset covering multiple standard planes. Extensive experiments on this in-house dataset, along with two public benchmarks, demonstrate that SEMC consistently outperforms state-of-the-art methods across all datasets.

References

Baumgartner, C. F.; Kamnitsas, K.; Matthew, J.; Fletcher, T. P.; Smith, S.; Koch, L. M.; Kainz, B.; and Rueckert, D. 2017. SonoNet: real-time detection and localisation of fetal standard scan planes in freehand ultrasound. *IEEE transactions on medical imaging*, 36(11): 2204–2215.

Baumgartner, C. F.; Kamnitsas, K.; Matthew, J.; Smith, S.; Kainz, B.; and Rueckert, D. 2016. Real-time standard scan plane detection and localisation in fetal ultrasound using fully convolutional neural networks. In *International conference on medical image computing and computer-assisted intervention*, 203–211. Springer.

Cai, Y.; Sharma, H.; Chatelain, P.; and Noble, J. A. 2018. Multi-task sonoeyenet: detection of fetal standardized planes assisted by generated sonographer attention maps. In *International conference on medical image computing and computer-assisted intervention*, 871–879. Springer.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A Simple Framework for Contrastive Learning of Visual Representations. arXiv:2002.05709.

Chen, X.; Zhou, Y.; Wu, D.; Yang, C.; Li, B.; Hu, Q.; and Wang, W. 2023. Area: adaptive reweighting via effective area for long-tailed classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 19277–19287.

Christodoulou, C. I.; Pattichis, C. S.; Pantziaris, M.; and Nicolaides, A. 2003. Texture-based classification of atherosclerotic carotid plaques. *IEEE transactions on medical imaging*, 22(7): 902–912.

Ciobanu, G.; Enache, I.-A.; Iovoaica-Rămescu, C.; Berbecaru, E. I. A.; Vochin, A.; Băluță, I. D.; Istrate-Ofițeru, A. M.; Comănescu, C. M.; Nagy, R. D.; Şerbănescu, M.-S.; et al. 2025. Automatic identification of fetal abdominal planes from ultrasound images based on deep learning. *Journal of Imaging Informatics in Medicine*, 1–8.

Di Cosmo, M.; Fiorentino, M. C.; Villani, F. P.; Frontoni, E.; Smerilli, G.; Filippucci, E.; and Moccia, S. 2022. A deep learning approach to median nerve evaluation in ultrasound images of carpal tunnel inlet. *Medical & Biological Engineering & Computing*, 60(11): 3255–3264.

Fedus, W.; Zoph, B.; and Shazeer, N. 2022. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. arXiv:2101.03961.

Guo, J.; Tan, G.; Wu, F.; Wen, H.; and Li, K. 2022. Fetal ultrasound standard plane detection with coarse-to-fine multi-task learning. *IEEE Journal of Biomedical and Health Informatics*, 27(10): 5023–5031.

Huang, X.; Chen, M.; Liu, P.; and Du, Y. 2020. Texture feature-based classification on transrectal ultrasound image for prostatic cancer detection. *Computational and Mathematical methods in Medicine*, 2020(1): 7359375.

Jin, Y.; Li, M.; Lu, Y.; Cheung, Y.-m.; and Wang, H. 2023. Long-tailed visual recognition via self-heterogeneous integration with knowledge excavation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 23695–23704.

Ke, T.-W.; Mo, S.; and Yu, S. X. 2024. Learning Hierarchical Image Segmentation For Recognition and By Recognition. arXiv:2210.00314.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2021. Supervised Contrastive Learning. arXiv:2004.11362.

Krishna, T. B.; and Kokil, P. 2024. Standard fetal ultrasound plane classification based on stacked ensemble of deep learning models. *Expert Systems with Applications*, 238: 122153.

Latha, S.; Samiappan, D.; and Kumar, R. 2020. Carotid artery ultrasound image analysis: A review of the literature. *Proceedings of the institution of mechanical engineers, Part H: journal of engineering in medicine*, 234(5): 417–443.

Leclerc, S.; Smistad, E.; Pedrosa, J.; Østvik, A.; Cervenansky, F.; Espinosa, F.; Espeland, T.; Berg, E. A. R.; Jodoin, P.-M.; Grenier, T.; et al. 2019. Deep learning for segmentation using an open large-scale dataset in 2D echocardiography. *IEEE transactions on medical imaging*, 38(9): 2198–2210.

Li, J.; Gao, Z.; Wang, C.; Pu, B.; and Li, K. 2025. A rule-guided interpretable lightweight framework for fetal standard ultrasound plane capture and biometric measurement. *Neurocomputing*, 621: 129290.

Liao, L.-J.; Cheng, P.-C.; and Chan, F.-T. 2024. Machine Learning on Ultrasound Texture Analysis Data for Characterizing of Salivary Glandular Tumors: A Feasibility Study. *Diagnostics*, 14(16): 1761.

Lin, H.; Yu, X.; Zhang, P.; Bai, X.; and Zheng, J. 2024. Consistent prototype contrastive learning for weakly supervised person search. *Journal of Visual Communication and Image Representation*, 105: 104321.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117–2125.

Lin, Z.; Li, S.; Ni, D.; Liao, Y.; Wen, H.; Du, J.; Chen, S.; Wang, T.; and Lei, B. 2019. Multi-task learning for quality assessment of fetal head ultrasound images. *Medical image analysis*, 58: 101548.

Maraci, M. A.; Napolitano, R.; Papageorghiou, A.; and Noble, J. A. 2014. Searching for structures of interest in an ultrasound video sequence. In *International Workshop on Machine Learning in Medical Imaging*, 133–140. Springer.

Men, Q.; Teng, C.; Drukker, L.; Papageorghiou, A. T.; and Noble, J. A. 2023. Gaze-probe joint guidance with multi-task learning in obstetric ultrasound scanning. *Medical image analysis*, 90: 102981.

Migliorelli, G.; Fiorentino, M. C.; Di Cosmo, M.; Villani, F. P.; Mancini, A.; and Moccia, S. 2024. On the use of contrastive learning for standard-plane classification in fetal ultrasound imaging. *Computers in Biology and Medicine*, 174: 108430.

Mildenberger, D.; Hager, P.; Rueckert, D.; and Menten, M. J. 2025. A Tale of Two Classes: Adapting Supervised Contrastive Learning to Binary Imbalanced Datasets. arXiv:2503.17024.

Prabakaran, B. S.; Hamelmann, P.; Ostrowski, E.; and Shafique, M. 2023. FPUS23: an ultrasound fetus phantom dataset with deep neural network evaluations for fetus orientations, fetal planes, and anatomical features. *IEEE Access*, 11: 58308–58317.

Pu, B.; Li, K.; Li, S.; and Zhu, N. 2021. Automatic fetal ultrasound standard plane recognition based on deep learning and IIoT. *IEEE Transactions on Industrial Informatics*, 17(11): 7771–7780.

Riquelme, C.; Puigcerver, J.; Mustafa, B.; Neumann, M.; Jenatton, R.; Susano Pinto, A.; Keysers, D.; and Houlsby, N. 2021. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems*, 34: 8583–8595.

Salomon, L.; Bernard, J.; Duyme, M.; Doris, B.; Mas, N.; and Ville, Y. 2006. Feasibility and reproducibility of an image-scoring method for quality control of fetal biometry in the second trimester. *Ultrasound in obstetrics & gynecology*, 27(1): 34–40.

Salomon, L. J.; Alfirevic, Z.; Berghella, V.; Bilardo, C.; Hernandez-Andrade, E.; Johnsen, S.; Kalache, K.; Leung, K.-Y.; Malinger, G.; Munoz, H.; et al. 2011. Practice guidelines for performance of the routine mid-trimester fetal ultrasound scan. *Ultrasound in Obstetrics & Gynecology*, 37(1).

Shazeer, N.; Mirhoseini, A.; Maziarz, K.; Davis, A.; Le, Q.; Hinton, G.; and Dean, J. 2017. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. arXiv:1701.06538.

Sowe, E. A. 2025. Momentum Contrast for Unsupervised Visual Representation Learning.

Spencer, J. K.; and Adler, R. S. 2008. Utility of portable ultrasound in a community in Ghana. *Journal of Ultrasound in Medicine*, 27(12): 1735–1743.

Wang, Y.; Yang, Q.; Drukker, L.; Papageorghiou, A.; Hu, Y.; and Noble, J. A. 2022. Task model-specific operator skill assessment in routine fetal ultrasound scanning. *International Journal of Computer Assisted Radiology and Surgery*, 17(8): 1437–1444.

Xie, H.; Wang, N.; He, M.; Zhang, L.; Cai, H.; Xian, J.; Lin, M.; Zheng, J.; and Yang, Y. 2020. Using deep-learning algorithms to classify fetal brain ultrasound images as normal or abnormal. *Ultrasound in Obstetrics & Gynecology*, 56(4): 579–587.

Yang, Y.; Fu, H.; Aviles-Rivero, A. I.; Schönlieb, C.-B.; and Zhu, L. 2023. DiffMIC: Dual-Guidance Diffusion Network for Medical Image Classification. arXiv:2303.10610.

Yu, T.; Tsui, P.-H.; Leonov, D.; Wu, S.; Bin, G.; and Zhou, Z. 2024. LPC-SonoNet: A Lightweight Network Based on SonoNet and Light Pyramid Convolution for Fetal Ultrasound Standard Plane Detection. *Sensors*, 24(23): 7510.

Yu, W.; Si, C.; Zhou, P.; Luo, M.; Zhou, Y.; Feng, J.; Yan, S.; and Wang, X. 2023. Metaformer baselines for vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(2): 896–912.

Zhu, H.; Salcudean, S.; and Rohling, R. 2022. Gaze-guided class activation mapping: Leverage human visual attention for network attention in chest x-rays classification. In *Proceedings of the 15th International Symposium on Visual Information Communication and Interaction*, 1–8.

Zhu, J.; Zhu, X.; Wang, W.; Wang, X.; Li, H.; Wang, X.; and Dai, J. 2022. Uni-Perceiver-MoE: Learning Sparse Generalist Models with Conditional MoEs. In Oh, A. H.; Agarwal, A.; Belgrave, D.; and Cho, K., eds., *Advances in Neural Information Processing Systems*.

Zoph, B. 2022. Designing effective sparse expert models. In *2022 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, 1044–1044. IEEE.