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

Qing Cai[1], GuihaoYan[1], Fan Zhang[2*], Cheng Zhang[1*], Zhi Liu[3]

[1]Faculty of Information Science and Engineering, Ocean University of China
[2]Department of Health Technology and Informatics, The Hong Kong Polytechnic University
[3]School of Information Science and Engineering, Shandong University
cq@ouc.edu.cn, {yanguihao, zhangcheng}@stu.ouc.edu.cn, fan-hti.zhang@polyu.edu.hk, liuzhi@sdu.edu.cn

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, ultimately resulting in suboptimal recognition of both structural and discriminative details in ultrasound standard planes. To address these issues, we propose SEMC, a novel Structure-Enhanced Mixture-of-Experts Contrastive learning framework that combines structure-aware feature fusion with expert-guided contrastive learning. Specifically, we first introduce a novel 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. Then, a novel 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 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.

Code — https://github.com/YanGuihao/SEMC
Datasets — https://github.com/YanGuihao/SEMC

Introduction

Ultrasound imaging is one of the most widely used medical imaging techniques in clinical practice, owing 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 and soft tissues and is widely used in prenatal examinations. In clinical workflows, acquiring standard planes (SP) is essential for accurate diagnosis and quantitative assessment. These planes provide clinicians with reliable structural visualization and consistent anatomical 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

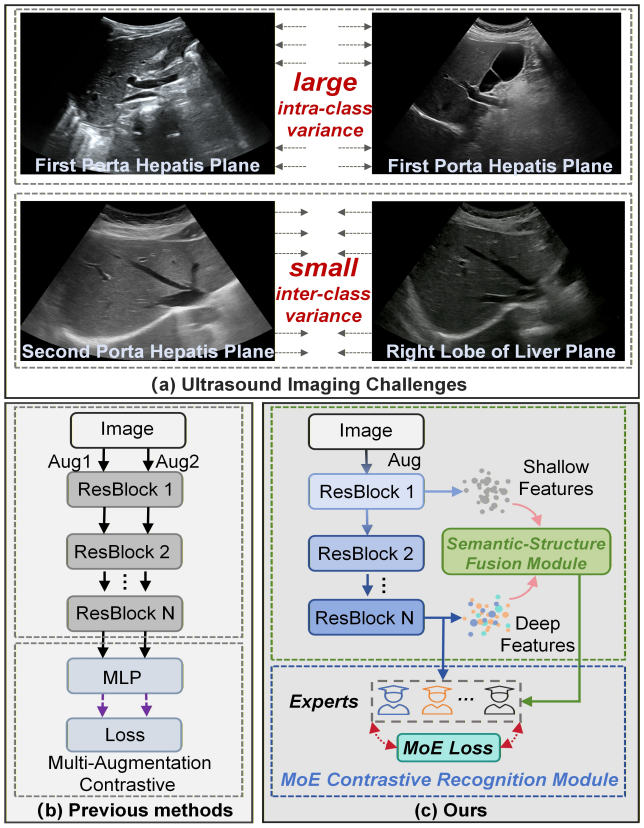

Figure 1: (a) Ultrasound standard planes exhibit large intra-class variance, where images from the same plane can appear markedly different, and small inter-class variance, where different planes often share highly similar visual patterns. (b) Previous methods mainly rely on deep semantic features, neglecting shallow structural cues. (c) In contrast, Our SEMC framework integrates the shallow structure via semantic-structure fusion and employs a MoE for hierarchical contrastive learning, which ca yields more discriminative and structure-aware representations.

*Corresponding authors.

length, head circumference, and abdominal circumference, respectively. These biometric measurements serve as key indicators for evaluating fetal growth (Salomon et al. 2006; Guo et al. 2022). However, the quality of acquired standard planes can vary considerably depending on the operator's experience and scanning technique, which may influence the accuracy of growth assessments and subsequent 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 face several critical challenges. The quality of ultrasound images varies considerably due to speckle noise, low contrast, and indistinct anatomical boundaries, making structural region detection inherently difficult. As shown in Figure 1(a), images from the same anatomical plane show substantial appearance variations caused by inconsistent acquisition angles, probe pressure, and operator experience (Lin et al. 2019; Xie et al. 2020; Yu et al. 2024; Krishna and Kokil 2024), while images from distinct planes often exhibit subtle visual differences due to low contrast and ambiguous boundaries, requiring fine-grained discrimination (Baumgartner et al. 2016, 2017). Most existing approaches focus primarily on deep semantic representations while overlooking shallow structural cues (Cai et al. 2018a; Zhang et al. 2024b; Yan et al. 2025), thereby limiting the model's ability in both semantic discrimination and structural perception (Men et al. 2023; Li et al. 2025; Liu, Ye, and Du 2024; Zhang et al. 2024a). Moreover, although contrastive learning has been incorporated as an auxiliary strategy by constructing augmented positive and negative pairs, these techniques often struggle to capture the fine-grained semantic distinctions inherent to ultrasound images, as illustrated in Figure 1(b).

In response to these gaps, we propose a novel structure-enhanced mixture-of-experts contrastive learning framework, dubbed SEMC, which effectively integrates structure-aware feature fusion with expert-guided contrastive learning to tackle the challenges inherent in ultrasound standard plane recognition. Specifically, in this framework, we design a novel Semantic-Structure Fusion Module (SSFM) that explicitly aligns and integrates shallow structural cues with deep semantic representations. It enhances the model's sensitivity to fine-grained structural details. To further enhance the model's discriminative capability, we designs a new 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. Additionally, we construct a high-quality liver ultrasound dataset, **LP2025**, containing six standard planes to address the scarcity of publicly available data for 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 introduce a novel structure-enhanced mixture-of-experts contrastive learning framework, dubbed SEMC, which integrates the semantic-structure fusion and MoE contrastive recognition modules to enhance fine-grained structural perception and discriminative feature representation for ultrasound plane recognition.
- We construct LP2025, a high-quality liver ultrasound dataset comprising six standard planes, addressing the scarcity of publicly available benchmarks and supporting further research in standard plane recognition.
- Extensive experiments on two public datasets and our in-house liver ultrasound dataset demonstrate that SEMC framework consistently outperforms existing state-of-the-art 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 and KNN), but their generalization capability was limited due to weak feature representation and low quality of ultrasound (Christodoulou et al. 2003; Latha, Samiappan, and Kumar 2020; Huang et al. 2020; Liao, Cheng, and Chan 2024). In recent years, convolutional neural networks (CNNs) have become the mainstream solution. For example, the SonoNet (Baumgartner et al. 2017) series, built on the VGG architecture, achieved promising performance in fetal standard plane recognition. Subsequent studies have incorporated multi-task learning, attention mechanisms (Cai et al. 2021; Zhang et al. 2024c), and structural priors to enhance the model's ability to identify key regions and capture fine-grained variations (Cai et al. 2018b; Zhu, Salcudean, and Rohling 2022; Yu et al. 2024; Ciobanu et al. 2025). However, these methods mainly rely on high-level features and often overlook shallow structural cues and spatial context, leading to degraded performance under subtle inter-class differences or complex imaging conditions. To overcome this, we introduce a semantic–structure fusion module that integrates shallow and deep features to enhance structural representation and discrimination.

Contrastive Learning and Mixture-of-Experts

Contrastive learning has shown strong potential for improving representation learning, particularly in medical imaging tasks where data are limited and class boundaries are ambiguous. Methods such as MoCo (He et al. 2020) and SimCLR (Chen et al. 2020) optimize the representation 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 enhances discriminability and semantic consistency by leveraging label information during pair con-

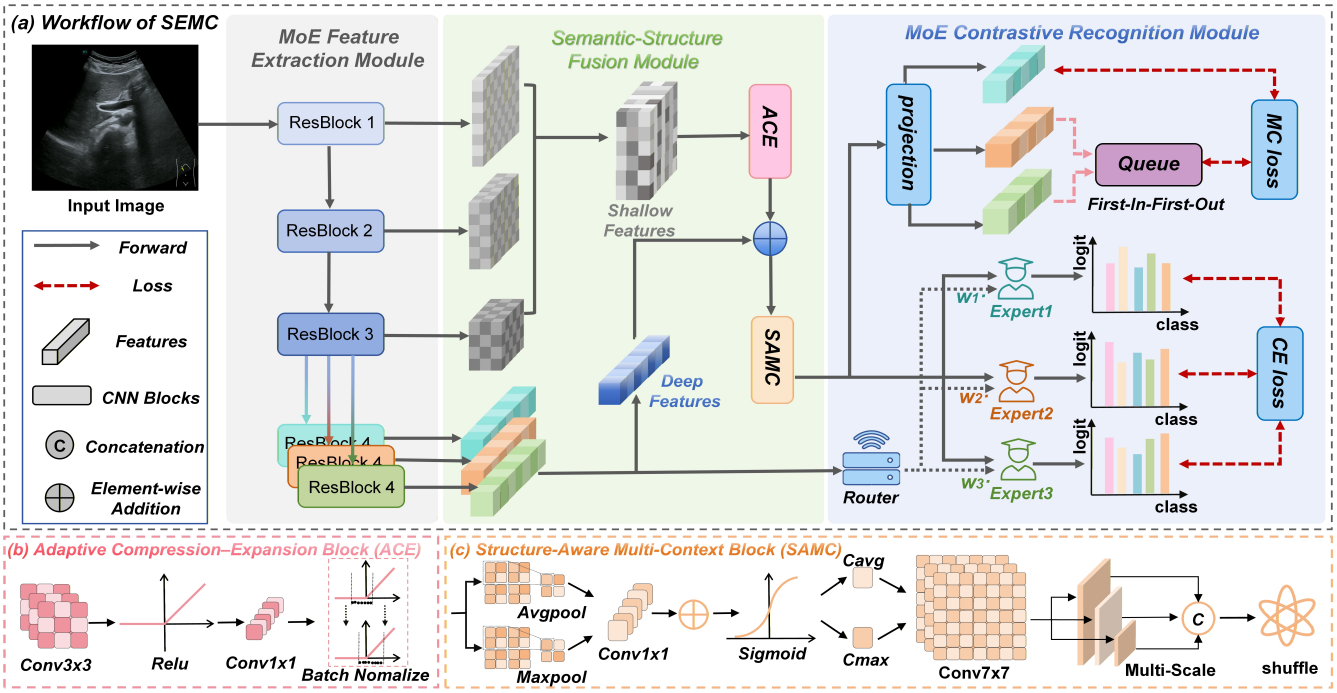

Figure 2: Architecture of the proposed SEMC framework. The framework first employs an MoE-based feature extractor to generate multi-level expert features from the input ultrasound image. These features are then aligned and enhanced through a Semantic-Structure Fusion Module (SSFM). The resulting representations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

struction. Meanwhile, the MoE paradigm has gained increasing attention for its dynamic modeling capabilities and parameter efficiency (Shazeer et al. 2017; Riquelme et al. 2021; Zoph 2022). For instance, Conditional MoE (Zhu et al. 2022) and Switch Transformer (Fedus, Zoph, and Shazeer 2022) have achieved substantial breakthroughs in both natural language processing and computer vision. Nevertheless, in ultrasound image analysis, where anatomical structures are complex and boundaries often indistinct, the integration of MoE with contrastive learning remains underexplored. Existing methods lack effective mechanisms to guide expert collaboration using structural cues. To this end, we propose a framework that combines structure-enhanced feature fusion with contrastive expert modeling,explicitly improving the model's ability to recognize fine-grained differences in standard plane classification tasks.

Methodology

Workflow Overview

We propose a novel Structure-Enhanced Mixture-of-Experts Contrastive (SEMC) learning framework for ultrasound standard plane recognition. As illustrated in Figure 2, our core contributions include two main components: (1) the semantic-structure fusion module, which explicitly aligns and integrates multi-level features to enhance structural awareness, and (2) the MoE-based contrastive recognition module, which 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. In the following sections, we will provide the details of these two modules as well as our in-house dataset.

LP2025 Dataset Construction

Data Collection. To advance research on standard plane recognition in ultrasound imaging, we introduce LP2025, a comprehensive and high-quality dataset specifically curated for deep learning-based anatomical understanding and classification. The dataset was developed under the clinical supervision of experienced radiologists and certified sonographers from a leading tertiary medical center. All ultrasound scans were acquired using high-resolution diagnostic systems to ensure excellent image quality and clear structural visibility. A standardized imaging protocol was strictly followed to harmonize scanning procedures across different patients and sessions, ensuring consistency in anatomical coverage, spatial resolution, and diagnostic relevance.

Each subject underwent a systematic abdominal ultrasound examination, during which six clinically meaningful liver standard planes were meticulously captured. These

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: Image distribution of the LP2025 dataset across six liver standard planes and the non-standard (NSP) category.

planes were selected based on their diagnostic relevance 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, comprising images that do not correspond to the above-defined diagnostic views but are frequently encountered in routine ultrasound examinations. The inclusion of this category introduces realistic variability and classification ambiguity, thereby enhancing model robustness and better reflecting 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 have more than 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 two sonographers to detect potential inconsistencies or errors.
- Consensus Verification: For cases with disagreement, at least three senior sonographers engaged in a consensus discussion to ensure clinically reliable labels.
- 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 could negatively influence model training or evaluation.

This comprehensive, multi-expert review process ensures the reliability and trustworthiness of the LP2025, establishing a robust foundation for both algorithm development and clinically oriented research.

Table 1 summarizes the LP2025 dataset, which contains 9,369 high-quality, clinically validated liver ultrasound images across six standard planes. All patient data were thoroughly anonymized, with no identifiable information retained during collection, processing, or release.

Mixture-of-Experts Feature Extraction Module

To capture fine-grained variations and complex anatomical structures in ultrasound images, we design a MoE feature extraction module. The first three blocks (*i.e.*, `layer1` to `layer3`) are shared across all branches and serve as a common encoder for extracting low- and mid-level features. Beyond `layer3`, we introduce three parallel, structurally identical yet parameter-independent fourth-stage blocks (denoted as `layer4-1`, `layer4-2`, and `layer4-3`), forming three specialized deep expert pathways:

$$F_1, F_2, F_3 = \text{ResNet}_{1\sim3}(x), \tag{1}$$

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

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

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

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

Semantic-Structure Fusion Module (SSFM)

Most existing methods primarily focus on deep features while overlooking the complementary value of shallow features, particularly in cases where anatomical contrast is weak or boundaries are indistinct. To address this limitation, we propose a novel Structure-Semantic Fusion Module (SSFM), which integrates shallow and deep features through two components, *i.e.*, 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 the spatial and channel mismatches between shallow and deep features, we propose a lightweight ACE module. It aligns the shallow features $\{F_1, F_2, F_3\}$ with the deep expert features through progressive downsampling and channel adaptation. Each ACE block processes an input $\mathbf{X}_0 = F_i$

through L sequential stages:

$$\mathbf{X}_{i+1} = \mathrm{BN}\Big(\mathrm{Conv}_{1\times1}\big(\mathrm{ReLU}\big(\mathrm{BN}\cdot \mathrm{DWConv}_{3\times3}^{s=2}(\mathbf{X}_i)\big)\big)\Big), \quad i = 0, 1, \ldots, L-1, \tag{5}$$

where the channel size doubles at each step, following $C_i = C_{\mathrm{in}} \times 2^i$. After L stages, a 1×1 convolution followed by BN and ReLU maps $\mathbf{X}_L$ to the target channel dimension C_{out}:

$$F_i' = \mathrm{ReLU}\left(\mathrm{BN}\left(\mathrm{Conv}_{1\times1}(\mathbf{X}_L)\right)\right), \tag{6}$$

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

ACE first reduces spatial resolution using strided depthwise convolutions, and then adjusts channels dimensions through pointwise convolutions, computational efficiency while preserving structural information. The aligned shallow features are subsequently fused with deep expert features $\{D_1, D_2, D_3\}$ through element-wise addition:

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

Compared with feature concatenation, this fusion strategy avoids channel redundancy, reduces parameters, and encourages the learning of shared discriminative patterns.

Structure-Aware Multi-Context Block (SAMC). To enhance the discriminative power and structural representation of the fused features, we propose a novel SAMC module, which reconstructs feature patterns across multiple receptive fields using a set of parallel multi-scale convolutions. Additionally, a coordinated channel–spatial attention mechanism adaptively highlights informative responses. By jointly modeling semantic cues and spatial structures, the SAMC enables model to capture fine-grained anatomical details while suppressing irrelevant background variations. For each fusion branch $\mathcal{M}_i$, the processing is performed 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 $\mathbf{C}_i \in \mathbb{R}^C$ is the adaptive channel attention. Global pooling captures context, while shared fully connected layers with activation $\delta(\cdot)$ capture inter-channel dependencies. The sigmoid $\sigma(\cdot)$ produces normalized attention weights.

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

where $\mathbf{S}_i \in \mathbb{R}^{1\times H\times W}$ denotes the spatial attention map generated from the channel-refined features. A convolutional layer aggregates spatial cues and guides the network to emphasize anatomically relevant regions.

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

where $\{\mathbf{F}_k^{(i)}\}_{k=1}^K$ denote the multi-scale features extracted from spatially enhanced input $\mathbf{S}_i \odot \mathbf{C}_i \odot \mathcal{M}_i$. These features are concatenated, channel-shuffled to facilitate cross-channel interaction, and compressed via pointwise convolution to produce 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 during training. However, these approaches often struggle to effectively capture the inherent fine-grained semantic variations in ultrasound images. To address this, we designs a novel MoE contrastive recognition module, which consists of two synergistic branches. (1) MoE Contrastive Branch: Multiple expert subnetworks focus on different regions of the feature space and collaboratively perform hierarchical contrastive learning. This enhances the inter-class separability and intra-class compactness in the learned representations. (2) MoE Recognition Branch: Expert subnetworks extract complementary discriminative cues from diverse semantic perspectives and spatial scales, thereby improving the model's ability to accurately recognize various standard planes.

MoE Contrastive Branch. To fully exploit multi-level semantic and spatial information for ultrasound plane recognition, we propose a MoE-Enhanced Contrastive Branch comprising three expert branches that share a common backbone but are supervised by different fusion views. The first expert output $\mathbf{O}_1$ is used as the contrastive anchor (query), whereas $\mathbf{O}_2$ and $\mathbf{O}_3$ serve as positive keys for updating a dynamic queue that supports negative sampling. Classification logits from all experts are concatenated, and the ground-truth labels are replicated for semantic supervision. The current expert features are further integrated with a momentum memory queue $\mathcal{Q}$, which stores historical representations to facilitate structural contrastive learning:

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

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

where $\mathbf{O}_i$ denote the expert outputs, and $\mathbf{y}$ and $\mathbf{y}_{\mathcal{Q}}$ are the labels for current batch and queue, respectively. Concatenating features along the batch dimension enlarges the pool of positive and negative pairs, improving contrastive learning.

We define two complementary losses: a supervised contrastive loss, which leverages label information to pull semantically similar samples closer together, and a self-supervised contrastive loss, which identifies positive pairs by mining class-consistent samples from both the current batch and the queue without relying on explicit labels:

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

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

The final objective is defined as a weighted sum of the two losses, controlled by a balancing factor λ:

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

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 feature patterns, but simple averaging may overlook sample-specific variations. To

Method	FPUS23				CAMUS			
	Accuracy↑	Precision↑	Recall↑	F1-score↑	Accuracy↑	Precision↑	Recall↑	F1-score↑
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
SEMC (Ours)	**95.78**	**94.38**	**95.81**	**95.06**	**82.13**	**82.03**	80.08	**80.93**

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

address this, we introduce a novel MoE classification branch equipped with a learnable sparse gating mechanism. Leveraging Gumbel-Softmax (Lin et al. 2017), the gate adaptively selects the most informative experts while remaining fully differentiable, improving robustness to anatomical ambiguity and imaging variability.

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

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

where τ controls the sparsity of the distribution.. Given expert predictions $\mathbf{z}_n \in \mathbb{R}^{B\times C}$, we stack them into $\mathbf{Z} \in \mathbb{R}^{B\times N\times C}$ 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 interaction, thus enhancing model generalization. To balance the main classification loss L_{moe} and the contrastive loss L_{mc}, we employ a lightweight adaptive network. Given an input feature $\mathbf{O} \in \mathbb{R}^{B\times C\times H\times W}$, it predicts a sample-specific weight $\alpha = g(\mathbf{O}) \in (0, 1)$. Overall, the total loss is computed as a weighted combination of the two losses:

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

The balancing factor α is adaptively adjusted based on sample difficulty and the dynamics of the training process. This mechanism removes the need for manually tuned hyperparameters and enables end-to-end learning of optimal weights. Consequently, it enhances the stability and effectiveness of multi-task collaborative training.

Experimental Results

Datasets and Evaluation Metrics

FPUS23 (Prabakaran et al. 2023) is a public fetal ultrasound dataset for standard plane recognition, covering key anatomical views such as the head, abdomen, femur, and thorax, with expert annotations suitable for supervised learning. **CAMUS** (Leclerc et al. 2019) is a cardiac ultrasound dataset originally designed for segmentation. we selected the apical two-chamber and four-chamber views and annotated them for classification. Its diversity across subjects makes it well suited for evaluating model generalization. Additionally, we use our in-house **LP2025** dataset, which contains clinically defined standard liver planes for abdominal ultrasound analysis. To fairly evaluate and compare our method with existing approaches, we adopt four commonly used metrics, including Accuracy, Precision, Recall, and F1-score.

Implementation Details

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

Comparison with State-of-the-Art Methods

To evaluate the classification performance of our model, we conducted a comparative analysis with recent 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 a fair comparison, all models are trained with consistent settings and evaluated under identical experimental conditions.

Results on FPUS23 Dataset. Table 2 shows that on the FPUS23 fetal standard plane dataset, our method achieves the highest performance across all metrics. Specifically, we obtain an Accuracy of 95.78%, surpassing the second-best MetaFormer (95.52%) by 0.26%. Our method also achieves the highest F1-score, reaching 95.06% and outperforming Area (94.40%) and SupMin (94.34%). These improvements

Method	LP2025			
	Accuracy↑	Precision↑	Recall↑	F1-score↑
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 comparisons of different models on the in-house LP2025 dataset.

ACE	SAMC	L_{mc}	Accuracy ↑	F1-score ↑
✗	✗	✗	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 the proposed ACE, SAMC, and L_{mc} on the LP2025 dataset.

stem from the proposed semantic-structure fusion module, which effectively captures both shallow structural cues and deep semantic features. Moreover, the cross-expert collaborative classification branch adaptively fuses predictions from multiple experts, thereby enhancing robustness and overall 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%). It demonstrate that SEMC addresses the large intra-class variability and high inter-class similarity characteristic of ultrasound imaging.

Results on LP2025 Dataset. Table 3 reports the results on the LP2025 dataset, where our method achieves the state-of-the-art performance. Specifically, it reaches an Accuracy of 82.30%, outperforming Diffmic (80.07%) by 2.23%, and obtains 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 superior classification performance.

Ablation Study

Ablation Study of Each Component. We conducted ablation studies on the LP2025 dataset to assess the contribution of each module in our framework. As shown in Table 4, we progressively incorporated the SSFM and MCRM into the baseline. The SSFM includes the ACE and SAMC submodules, while MCRM introduces a contrastive loss, *i.e.*, L_{mc}, to encourage expert branches to learn complementary representations. The baseline achieves 80.26% accuracy and 76.98% F1-score. Adding ACE increases accuracy to 81.38% and F1-score to 77.82%, demonstrating its effectiveness in capturing shallow structural cues. Besides, adding SAMC on top of ACE further enhances 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, confirming that the three components are complementary and jointly enhance both feature representation and generalization.

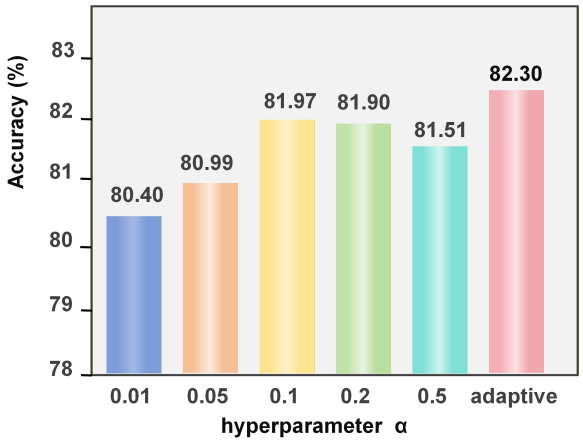

Figure 3: Performance comparison for different values of the hyperparameter α on the LP2025 dataset.

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

This paper presents a novel structure-enhanced mixture-of-experts contrastive learning framework, dubbed SEMC, for ultrasound standard plane recognition. Specifically, the proposed semantic-structure fusion module aligns and integrates shallow structural cues with deep semantic representations. Additionally, the mixture-of-experts contrastive recognition module is designed to specialize in different aspects of the feature space and collaboratively performs hierarchical contrastive learning, enabling the capture of fine-grained discriminative representations. More importantly, we establish a high-quality ultrasound dataset comprising six standard planes, addressing the scarcity of publicly available benchmarks. Extensive experiments on this in-house dataset and two public benchmarks demonstrate that SEMC consistently outperforms recent state-of-the-art methods.

Acknowledgments

This work was supported in part by the National Science Foundation of China under Grant62471448; in part by Shandong Provincial Natural Science Foundation under Grant ZR2024YQ004; in part by TaiShan Scholars Youth Expert Program of Shandong Province under Grant No.tsqn202312109.

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, Q.; Liu, H.; Zhou, S.; Sun, J.; and Li, J. 2018a. An adaptive-scale active contour model for inhomogeneous image segmentation and bias field estimation. *Pattern Recognition*, 82: 79–93.

Cai, Q.; Qian, Y.; Zhou, S.; Li, J.; Yang, Y.-H.; Wu, F.; and Zhang, D. 2021. AVLSM: Adaptive variational level set model for image segmentation in the presence of severe intensity inhomogeneity and high noise. *IEEE Transactions on Image Processing*, 31: 43–57.

Cai, Y.; Sharma, H.; Chatelain, P.; and Noble, J. A. 2018b. 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, Ş. G.; Enache, I.-A.; Iovoaica-Rămescu, C.; Berbecaru, E. I. A.; Vochin, A.; Băluţă, I. D.; Istrate-Ofiţeru, 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.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9729–9738.

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.

Liu, F.; Ye, M.; and Du, B. 2024. Learning a generalizable re-identification model from unlabelled data with domain-agnostic expert. *Visual Intelligence*, 2.

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.

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.

Yan, K.; Cai, Q.; Zhang, F.; Cao, Z.; and Liu, Z. 2025. SGTC: Semantic-guided triplet co-training for sparsely annotated semi-supervised medical image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39(9), 9112–9120.

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.

Zhang, F.; Liu, H.; Cai, Q.; Feng, C.-M.; Wang, B.; Wang, S.; Dong, J.; and Zhang, D. 2024a. Federated cross-incremental self-supervised learning for medical image segmentation. *IEEE Transactions on Neural Networks and Learning Systems*.

Zhang, F.; Liu, H.; Wang, J.; Lyu, J.; Cai, Q.; Li, H.; Dong, J.; and Zhang, D. 2024b. Cross co-teaching for semi-supervised medical image segmentation. *Pattern Recognition*, 152: 110426.

Zhang, T.; Cai, Q.; Gao, F.; Qi, L.; and Dong, J. 2024c. Exploring cross-domain few-shot classification via frequency-aware prompting. *arXiv preprint arXiv:2406.16422*.

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.