SHE-NET: SEMANTIC-GUIDED HIERARCHICAL EXPERTS FOR ULTRASOUND STANDARD PLANE RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Ultrasound standard plane recognition is an essential yet challenging task in clinical screening and diagnosis, owing to substantial intra-class variations and inter-class similarities caused by imaging noise, operator dependency, and anatomical variability. Existing methods mainly rely on visual appearance modeling, which may struggle to capture subtle structural differences and often overlook explicit anatomical semantics relevant to standard plane recognition. To address these challenges, we propose **SHE-Net**, a Semantic-guided Hierarchical Experts Network for ultrasound standard plane recognition. We first construct multi-granularity textual descriptions for existing datasets, providing explicit anatomical semantics from coarse plane categories to fine-grained structural characteristics. Based on these descriptions, we develop a Hierarchical Vision-Language Alignment **(HVLA)** strategy to align visual representations at different hierarchical levels with corresponding textual semantics, introducing structured semantic supervision into visual representation learning. Furthermore, we propose a Semantic-Routed Anatomical Experts **(SRAE)** module, which uses semantic-aware routing to dynamically assign visual representations to specialized experts, enabling the learning of complementary anatomical patterns from local structures to global plane configurations. By jointly integrating hierarchical semantic alignment and adaptive expert specialization, SHE-Net enhances the discriminability and structural consistency of ultrasound standard plane representations. In addition, we construct and annotate a new kidney ultrasound standard plane dataset to evaluate the generalizability of the proposed framework. Extensive experiments on three ultrasound standard plane datasets demonstrate that SHE-Net consistently outperforms existing methods across multiple evaluation metrics, highlighting the effectiveness of hierarchical anatomical semantics and semantic-guided expert specialization for robust ultrasound standard plane recognition.

1 INTRODUCTION

Ultrasound imaging is widely used in clinical practice due to its noninvasive nature, real-time capability, efficiency, and cost-effectiveness (Elangovan & Jeyaseelan, 2016; Kremkau, 2015; Cootney, 2001). In clinical workflows, standard planes provide reliable anatomical views and consistent reference points for accurate measurement and diagnosis (Spencer & Adler, 2008; Wang et al., 2022a). However, their acquisition is highly dependent on operator expertise and scanning techniques, leading to substantial variations in image quality and consistency (Moore & Copel, 2011; Salomon et al., 2011). Therefore, ultrasound standard plane recognition has become an important task for automatically identifying diagnostically meaningful views, reducing reliance on operator experience and enabling more consistent anatomical assessment and quantitative analysis (Di Cosmo et al., 2022; Xie et al., 2020; Maraci et al., 2014; Guo et al., 2022). Recent studies have explored deep learning-based methods for ultrasound standard plane recognition and achieved promising performance (Pu et al., 2021; Migliorelli et al., 2024). However, the task remains challenging due to substantial intra-class variations and inter-class similarities caused by imaging noise, operator dependency, and anatomical variability. Existing methods mainly rely on visual appearance modeling, which may struggle to distinguish subtle structural differences and often overlook explicit anatomical semantics essential for

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

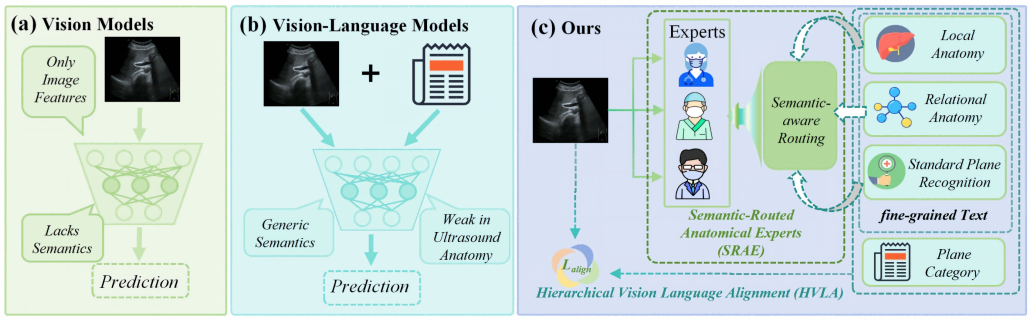


Figure 1: Comparison of different modeling paradigms for ultrasound standard plane recognition. (a) Traditional vision-only models lack semantic guidance. (b) Generic vision-language models fail to capture specific ultrasound anatomy. (c) Our proposed SHE-Net employs a hierarchical multi-expert architecture guided by medical semantics to mimic clinical reasoning.

accurate recognition (Baumgartner et al., 2016; Yu et al., 2024; Krishna & Kokil, 2024). As illustrated in Fig. 1(a), vision-only models learn discriminative patterns from image appearance without explicit semantic guidance, limiting their ability to capture clinically meaningful anatomical cues. In clinical practice, clinicians recognize standard planes through hierarchical anatomical reasoning, including identifying local structures, understanding inter-organ spatial relationships, and verifying global plane configurations. Such reasoning helps distinguish visually similar planes by integrating anatomical cues at different levels of abstraction. This motivates the incorporation of structured anatomical semantics into standard plane recognition. Recent vision-language approaches have introduced textual information to enhance visual representation learning, but most are designed for general-purpose vision-language tasks and may not adequately capture domain-specific anatomical knowledge (Bannur et al., 2023; Jiang et al., 2026; Li et al., 2024), as illustrated in Fig. 1(b). Moreover, they often lack fine-grained and hierarchical semantic guidance, making it difficult to establish effective correspondence between visual representations and clinically relevant anatomical concepts. Consequently, existing methods (Baumgartner et al., 2016; Yang et al., 2023; Cai et al., 2026) remain limited in providing semantically interpretable representations, particularly for challenging cases with subtle anatomical variations and high inter-class similarity. These limitations motivate semantic-guided mechanisms that incorporate hierarchical anatomical knowledge into visual representation learning for more reliable ultrasound standard plane recognition.

To address these limitations, we propose **SHE-Net**, a semantic-guided hierarchical experts framework tailored for ultrasound standard plane recognition. As illustrated in Fig. 1(c), SHE-Net explicitly incorporates structured anatomical semantics into visual representation learning, bridging the gap between appearance-based visual modeling and hierarchical anatomical reasoning. Specifically, we construct multi-granularity textual descriptions covering coarse anatomical configurations and fine-grained structural characteristics. Based on these descriptions, we develop a Hierarchical Vision-Language Alignment (**HVLA**) strategy, which provides structured semantic supervision by aligning visual representations with textual semantics at different levels of granularity. Furthermore, we introduce the Semantic-Routed Anatomical Experts (**SRAE**) module, which employs semantic-aware routing to dynamically activate specialized experts according to the anatomical semantics of the input. By encouraging different experts to focus on complementary anatomical patterns, SRAE facilitates the modeling of discriminative structures and their spatial configurations, thereby improving the consistency and robustness of standard plane recognition. Beyond the model design, the progress of ultrasound standard plane recognition is also hindered by the limited availability of publicly accessible datasets, particularly for specific organs such as the kidney. To address this gap, we collect and annotate a new kidney ultrasound standard plane dataset, termed **KU2026**, providing a dedicated benchmark for kidney standard plane recognition and facilitating further research in this underexplored domain. Extensive experiments on three ultrasound standard plane datasets demonstrate the effectiveness and generalizability of SHE-Net, achieving highly competitive performance across the evaluated benchmarks, with the best performance on most evaluation metrics.

We summarize our contributions as follows:

- We propose **SHE-Net**, a semantic-guided hierarchical experts framework for ultrasound standard plane recognition, which explicitly incorporates structured anatomical semantics into visual representation learning to alleviate visual ambiguities caused by intra-class variations and inter-class similarities.
- We develop a **Hierarchical Vision-Language Alignment (HVLA)** strategy based on multi-granularity textual descriptions, providing structured semantic supervision at different levels of abstraction and enhancing the discriminability of visual representations.
- We introduce the **Semantic-Routed Anatomical Experts (SRAE)** module, which employs semantic-aware routing to dynamically activate specialized experts, enabling complementary modeling of diverse anatomical patterns and improving the robustness of standard plane recognition.
- We establish **KU2026**, a new kidney ultrasound standard plane dataset that expands the availability of benchmarks for kidney standard plane recognition. Extensive experiments across three datasets demonstrate the effectiveness and generalizability of SHE-Net, with competitive or state-of-the-art performance across the evaluated benchmarks.

2 Related Work

2.1 Ultrasound Standard Plane Recognition

Ultrasound standard plane recognition is typically formulated as a visual classification or detection task for identifying clinically meaningful views. Early studies mainly relied on handcrafted features and conventional machine learning, while deep learning established CNNs as the dominant paradigm. SonoNet (Baumgartner et al., 2016) introduced CNN-based real-time detection and localization of fetal standard planes, followed by extensions to specific anatomical regions and standard plane recognition (Yu et al., 2016; Yaqub et al., 2017; Guo et al., 2022). Subsequent studies incorporated attention and anatomical information to enhance visual representation, including attention-gated networks (Schlemper et al., 2019), gaze-based learning (Droste et al., 2019), multi-scale feature modeling (Yu et al., 2024; Cai et al., 2026), and knowledge-guided anatomical priors (Zhao et al., 2022). More recently, transformer-based and hybrid architectures have been adopted to capture global dependencies and anatomical context. Vision Transformer-based methods (Tehsin et al., 2025) exploit global self-attention, while CNN-transformer hybrids (Chen et al., 2024) combine local and global feature modeling. Despite these advances, existing methods remain predominantly appearance-driven, with anatomical information typically serving as auxiliary cues or structural constraints. They lack explicit multi-granularity semantic modeling, limiting their ability to capture subtle structural differences and hierarchical anatomical relationships, particularly among visually similar standard planes.

2.2 Semantic Modeling in Medical Imaging

Semantic knowledge has been increasingly incorporated into visual representation learning in medical imaging. CLIP (Radford et al., 2021) established a general vision-language paradigm through large-scale image-text contrastive learning, which has been adapted to the medical domain by methods such as ConVIRT (Zhang et al., 2022) and MedCLIP (Wang et al., 2022b) using paired or unpaired medical image-text data. Subsequent studies have explored finer-grained semantic modeling, including global-local alignment in GLoRIA (Huang et al., 2021), biomedical knowledge integration in BioViL (Bannur et al., 2023), and fine-grained cross-modal and anatomical semantic alignment (Jiang et al., 2026; Li et al., 2024). Despite these advances, existing approaches mainly focus on general medical vision-language learning and rarely model the hierarchical structure of anatomical semantics. They lack explicit multi-granularity supervision that connects visual representations with anatomical concepts at different levels, limiting their ability to capture hierarchical anatomical relationships in ultrasound standard plane recognition.

2.3 Dynamic Expert Learning and Routing

Dynamic expert learning, typically formulated as mixture-of-experts (MoE), improves model capacity by combining specialized experts with input-dependent routing. Early sparse MoE mod-

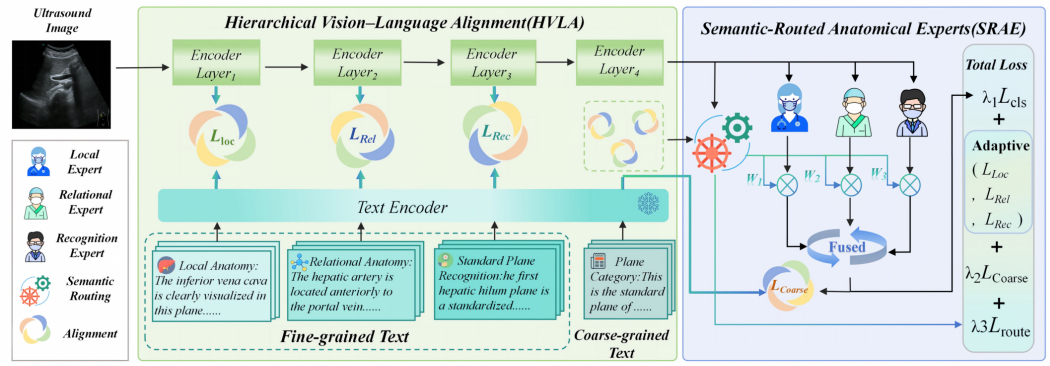


Figure 2: Overview of the proposed framework. It integrates **Hierarchical Vision-Language Alignment (HVLA)** for hierarchical semantic supervision, and **Semantic-Routed Anatomical Experts (SRAE)** for semantic-guided routing among specialized Local, Relational, and Recognition experts.

els (Shazeer et al., 2017; Lepikhin et al., 2020; Fedus et al., 2022) established scalable conditional computation, while subsequent vision MoE methods (Riquelme et al., 2021; Fan et al., 2022; Han et al., 2025) explored input-dependent expert selection and flexible routing strategies. Differentiable routing (Hazimeh et al., 2021) further enables adaptive expert assignment, and medical imaging studies have adopted expert-based architectures to model heterogeneous anatomical patterns (Liu et al., 2024). Despite these advances, existing expert frameworks mainly rely on visual features or learned routing scores for expert selection, without explicitly incorporating structured semantic knowledge. Consequently, expert specialization remains largely data-driven, lacking clear correspondence with anatomical concepts and limiting semantic-guided expert assignment for ultrasound standard plane recognition.

3 METHODS

3.1 OVERVIEW

We propose a **Semantic-guided Hierarchical Experts (SHE-Net)** framework for ultrasound standard plane recognition. As illustrated in Figure 2, SHE-Net consists of two complementary components: (1) the **Hierarchical Vision–Language Alignment (HVLA)** module, which aligns hierarchical visual features with structured textual semantics, and (2) the **Semantic-Routed Anatomical Experts (SRAE)** module, which learns complementary anatomical representations through specialized experts and adaptive visual routing.

Specifically, HVLA provides semantic supervision at local anatomy, relational anatomy, and standard plane recognition levels, while SRAE employs three corresponding experts with dynamically determined contributions. Textual information is used only during training, enabling SHE-Net to learn semantically guided and complementary anatomical representations for standard plane recognition.

3.2 HIERARCHICAL VISION-LANGUAGE ALIGNMENT

The **Hierarchical Vision-Language Alignment (HVLA)** module introduces structured medical semantics into visual representation learning by aligning hierarchical visual features with category-specific textual descriptions. Rather than imposing a single semantic constraint on the final representation, HVLA provides progressively higher-level supervision from local anatomy to relational anatomy and standard plane recognition.

Given an input ultrasound image x, the shared backbone produces hierarchical feature maps:

$$F_1, F_2, F_3, F_4 = \text{Backbone}(x). \tag{1}$$

HVLA operates on the first three levels, while F_4 is directly forwarded to the subsequent **Semantic-Routed Anatomical Experts (SRAE)** module. For $l \in 1, 2, 3$, the visual feature is projected into a

shared semantic space as

$$v_l = \mathrm{Norm}\left(\phi_l(\mathrm{GAP}(F_l))\right), \tag{2}$$

where $\phi_l(\cdot)$ denotes a level-specific learnable projection. These three visual levels are respectively associated with *local anatomy*, *relational anatomy*, and *standard plane recognition*. For a training sample with ground-truth class y, a textual embedding is randomly sampled from the corresponding class-specific semantic pool:

$$t_l \sim \mathcal{T}_y^l, \quad l \in \mathrm{loc}, \mathrm{rel}, \mathrm{rec}, \tag{3}$$

where each pool contains multiple descriptions of the same semantic level.

The visual and textual representations are then aligned at each semantic level using a prototype-based visual–semantic alignment objective:

$$\mathcal{L}_l = \mathrm{Align}(v_l, t_l, y), \quad l \in \mathrm{loc}, \mathrm{rel}, \mathrm{rec}. \tag{4}$$

Consequently, $\mathcal{L}_{\mathrm{loc}}$, $\mathcal{L}_{\mathrm{rel}}$, and $\mathcal{L}_{\mathrm{rec}}$ impose complementary semantic constraints on increasingly abstract visual representations. These hierarchical alignment signals are subsequently used to guide expert routing in SRAE, while the deepest feature F_4 is reserved for expert-based representation learning.

3.3 Semantic-Routed Anatomical Experts

The **Semantic-Routed Anatomical Experts (SRAE)** module learns complementary anatomical representations from the shared high-level feature F_4. It utilizes three functionally specialized experts to process the same feature, where an input-dependent gating network dynamically modulates their respective contributions.

Given $F_4 \in \mathbb{R}^{C\times H\times W}$ with $C = 2048$, three experts are constructed as

$$E_{\mathrm{loc}}(F_4), \quad E_{\mathrm{rel}}(F_4), \quad E_{\mathrm{rec}}(F_4), \tag{5}$$

where all experts output 512-dimensional representations. The Local Expert adopts stacked 3×3 convolutions to highlight fine-grained anatomical structures. The Relational Expert leverages ASPP to capture multi-scale anatomical relationships, and the Recognition Expert employs self-attention to model global configurations essential for standard plane recognition.

Expert weights are dynamically predicted from F_4 by the gating network:

$$\mathbf{w} = \mathrm{Softmax}(G(F_4)) = [w_{\mathrm{loc}}, w_{\mathrm{rel}}, w_{\mathrm{rec}}]^\top, \tag{6}$$

and the fused feature is obtained via weighted aggregation:

$$F_{\mathrm{fused}} = w_{\mathrm{loc}}E_{\mathrm{loc}}(F_4) + w_{\mathrm{rel}}E_{\mathrm{rel}}(F_4) + w_{\mathrm{rec}}E_{\mathrm{rec}}(F_4). \tag{7}$$

Since the gating network takes only visual features as input, textual information is not required during inference.

To transfer the hierarchical semantic knowledge learned by HVLA into the visual routing mechanism, we derive a semantic routing signal from the hierarchical semantic scores:

$$\mathbf{s} = [s_{\mathrm{loc}}, s_{\mathrm{rel}}, s_{\mathrm{rec}}], \tag{8}$$

where s_{loc}, s_{rel}, and s_{rec} measure the semantic consistency between the visual representation and the corresponding local, relational, and recognition-level semantics, respectively. To obtain a stable semantic routing target, we maintain an exponential moving average of the hierarchical semantic scores:

$$\mathbf{m}^{(t)} = \mu\mathbf{m}^{(t-1)} + (1-\mu)\mathbf{s}^{(t)}, \tag{9}$$

where $\mathbf{m}^{(t)}$ denotes the running semantic score state at training step t, and μ is the momentum coefficient. The smoothed semantic scores are then converted into a routing target distribution using a temperature-scaled softmax:

$$\mathbf{q} = \mathrm{Softmax}\left(\frac{\mathbf{m}^{(t)}}{\tau_r}\right), \tag{10}$$

where τ_r controls the sharpness of the routing target. The resulting semantic target distribution supervises the visual gate through the KL-divergence loss:

$$\mathcal{L}_{\mathrm{route}} = D_{\mathrm{KL}}\left(\mathbf{q} \,\|\, \mathbf{w}\right), \tag{11}$$

thereby transferring the hierarchical semantic knowledge learned during training to the visual routing mechanism without requiring textual inputs at inference time.

3.4 Loss Function

The overall objective jointly optimizes hierarchical semantic alignment, semantic-guided expert routing, and standard plane classification. In addition to the three hierarchical alignment losses, we impose a coarse semantic constraint on the fused representation:

$$v_{\text{coarse}} = \text{Norm}\left(\text{GAP}(F_{\text{fused}})\right), \qquad \mathcal{L}_{\text{coarse}} = \text{Align}\left(v_{\text{coarse}}, t_{\text{coarse}}, y\right), \tag{12}$$

where t_{coarse} is sampled from the category-specific coarse semantic pool. This constraint encourages the fused representation to preserve category-level semantic information.

To adaptively emphasize semantic levels with larger alignment losses, we compute

$$\boldsymbol{\alpha} = \text{Softmax}\left([\mathcal{L}_{\text{loc}}, \mathcal{L}_{\text{rel}}, \mathcal{L}_{\text{rec}}]^\top\right), \tag{13}$$

and obtain

$$\mathcal{L}_{\text{align}} = \alpha_{\text{loc}}\mathcal{L}_{\text{loc}} + \alpha_{\text{rel}}\mathcal{L}_{\text{rel}} + \alpha_{\text{rec}}\mathcal{L}_{\text{rec}} + 0.5\mathcal{L}_{\text{coarse}}. \tag{14}$$

Here, larger alignment losses indicate semantic levels that are relatively under-optimized and therefore receive stronger supervision.

Finally, the overall training objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{align}} + \lambda_{\text{route}}\mathcal{L}_{\text{route}}, \qquad \lambda_{\text{route}} = 0.2, \tag{15}$$

where $\mathcal{L}_{\text{cls}}$ is the cross-entropy loss for standard plane recognition, and $\mathcal{L}_{\text{route}}$ is the semantic score-based routing constraint introduced in Section 3.3.

4 Dataset Collection

To support research on kidney standard plane recognition in ultrasound imaging, we construct a new dataset, termed **KU2026**. Publicly available datasets for ultrasound standard plane recognition are predominantly focused on fetal and cardiac imaging, while dedicated benchmarks for renal ultrasound remain limited. This lack of data restricts the development and systematic evaluation of learning-based methods for kidney standard plane recognition.

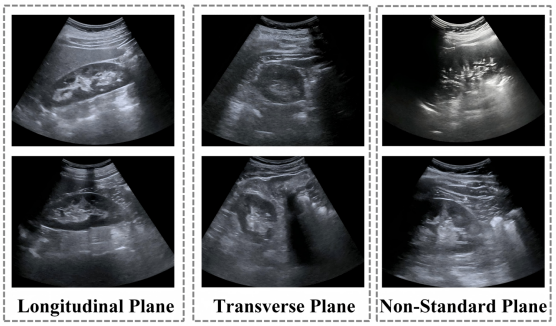


Figure 3: Illustrative examples of standard and non-standard renal ultrasound views in the KU2026 dataset.

KU2026 was collected in a clinical setting under the supervision of professional radiologists and experienced sonographers. The images were acquired during routine renal ultrasound examinations using clinical ultrasound systems, preserving the anatomical and imaging variations encountered in real-world practice,as illustrated in Fig. 3.

The KU2026 dataset contains three categories: **Transverse Plane (TP)**, **Longitudinal Plane (LP)**, and **Non-Standard Plane (NSP)**, with the class distribution summarized in Table 1. TP represents the axial view of the kidney, where the renal sinus and surrounding parenchyma provide structural cues for transverse anatomical assessment, while LP represents the longitudinal view, providing cues for renal length, parenchymal structure, and overall morphology. NSP includes images that do not satisfy the predefined criteria for either standard view, such as oblique and other non-canonical views commonly encountered in routine examinations. All images were independently reviewed and cross-validated by trained sonographers experienced in renal ultrasound, with annotation focused on standard plane categorization and the visibility of relevant anatomical structures.

Table 1: Statistics of the KU2026 dataset.

Plane	Abbreviation	Number
Transverse	TP	960
Longitudinal	LP	1384
Non-Standard	NSP	1372
Total	–	3716

Table 2: Quantitative comparison of different models on the FPUS23, LP2025, and KU2026 datasets. Best results are highlighted in **bold**.

	FPUS23				LP2025				KU2026			
Method	Acc↑	Pre↑	Rec↑	F1↑	Acc↑	Pre↑	Rec↑	F1↑	Acc↑	Pre↑	Rec↑	F1↑
Area	95.20	93.37	95.71	94.40	80.39	75.43	76.21	75.04	83.02	82.71	83.24	82.88
Shike	95.15	93.80	94.93	94.31	80.26	75.76	77.63	76.19	82.81	82.38	83.21	82.72
Cast	95.24	94.23	93.99	94.43	80.86	74.87	79.59	77.00	83.55	83.27	83.98	83.62
USFM	94.93	94.13	93.61	93.86	79.21	75.53	76.52	75.61	81.67	81.40	82.14	81.68
ALTA	94.03	92.20	94.17	93.08	78.95	72.36	74.21	72.40	82.21	82.20	82.53	82.35
TA-CST	95.25	93.77	95.54	94.57	82.29	73.29	75.22	73.79	83.29	82.87	83.59	83.07
IVAAN	95.82	94.21	94.65	94.33	82.97	77.27	79.81	78.42	83.99	82.84	83.38	82.85
Ours	**95.98**	**94.23**	**96.34**	**95.20**	**84.68**	**78.50**	**81.08**	**79.36**	**84.52**	**83.76**	**84.23**	**84.13**

Prior to dataset construction, all images were fully anonymized to remove patient-identifiable information. The resulting KU2026 dataset contains **3,716** renal ultrasound images and provides a dedicated benchmark for kidney standard plane recognition while preserving the variability of routine ultrasound acquisition.

5 Experiments

5.1 Experimental Setups

Datasets. We evaluate our method on three ultrasound standard plane recognition datasets. FPUS23 (Prabakaran et al., 2023) is a public fetal ultrasound dataset covering standard planes of the head, abdomen, femur, and thorax, with expert annotations. LP2025 is a liver ultrasound dataset containing six clinically defined standard planes and one non-standard plane category, covering diverse anatomical structures and visual appearances. KU2026 is a newly collected renal ultrasound dataset with standard plane annotations, providing an additional evaluation across different organs and imaging conditions.

Evaluation Metrics. We use four widely adopted metrics, including Accuracy, Precision, Recall, and F1-score. Accuracy measures the overall classification performance, while Precision and Recall evaluate the correctness and completeness of predictions, respectively. To account for class imbalance, Precision, Recall, and F1-score are computed using macro-averaging, assigning equal importance to all classes.

Implementation Details. Our method is implemented in PyTorch and trained on a single NVIDIA RTX 3090 GPU with Python 3.8. All input images are resized to 512×512. During training, we apply random rotation, horizontal and vertical flipping, and brightness perturbation for data augmentation. We use stochastic gradient descent (SGD) with a momentum of 0.9 and a weight decay of 1×10^{-4}. The initial learning rate is 1×10^{-3} and is decayed using cosine annealing. The model is trained for 200 epochs with a batch size of 32.

5.2 Comparison with State-of-the-Art

Quantitative Comparison. As shown in Table 2, our method achieves the highest Accuracy and F1-score across all three datasets. On FPUS23, it obtains 95.98% Accuracy and 95.20% F1-score, improving over the strongest competing methods by 0.16 and 0.63 percentage points, respectively (Chen et al., 2023). On LP2025, our method achieves 84.68% Accuracy and 79.36% F1-score, surpassing IVAAN (Jeong et al., 2026) by 1.71 and 0.94 percentage points, respectively. On KU2026, it achieves 84.52% Accuracy and 84.13% F1-score, exceeding the strongest competing results by 0.53 and 0.51 percentage points, respectively. Notably, the vision-language methods (Jiao et al., 2024), including ALTA (Lian et al., 2025), TA-CST (Nandam et al., 2025), and IVAAN, do not consistently outperform the visual baselines. This suggests that simply introducing cross-modal representations may not be sufficient for fine-grained ultrasound standard plane recognition. In con-

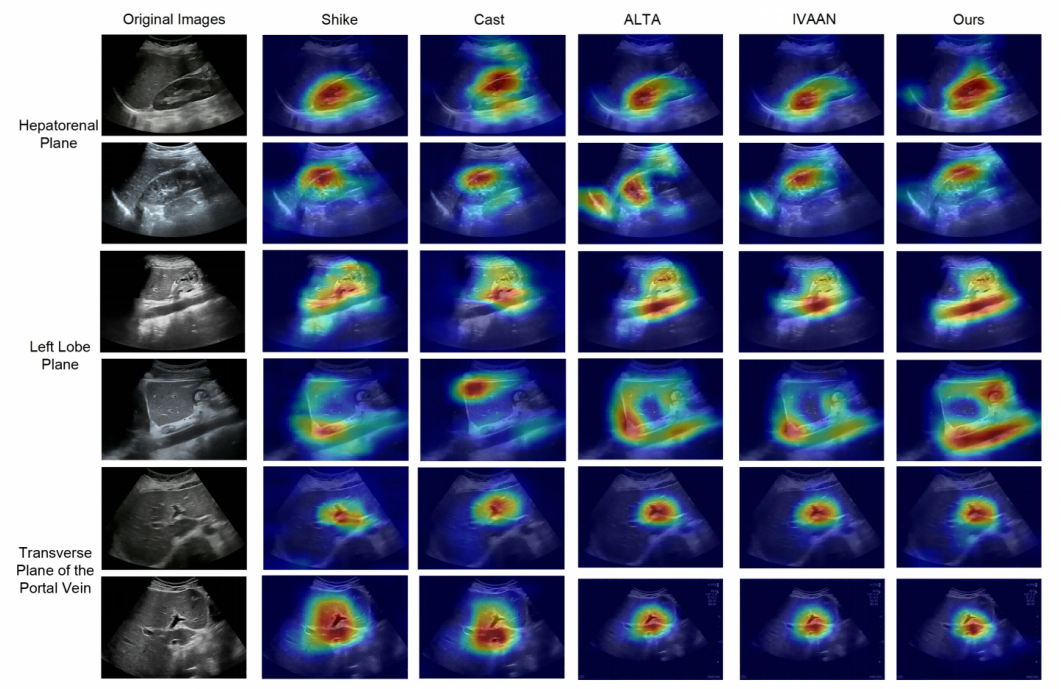


Figure 4: Visual comparison of Grad-CAM heatmaps produced by different methods on three representative ultrasound standard planes: (Top) Hepatorenal Plane, (Middle) Left Lobe Plane, and (Bottom) Transverse Plane of the Portal Vein. From left to right: Original image, Shike, Cast, ALTA, IVAAN, and our method.

trast, SHE-Net explicitly incorporates hierarchical anatomical semantics into visual representation learning, capturing complementary cues from local structures, anatomical relationships, and plane-specific characteristics.

Qualitative Comparison. Figure 4 presents Grad-CAM comparisons with four representative baselines, including Shike (Jin et al., 2023), Cast (Ke et al., 2024), ALTA, and IVAAN, on three clinically important planes: the Hepatorenal Plane, Left Lobe Plane, and Transverse Plane of the Portal Vein. Compared with the baselines, our method produces more focused and anatomically meaningful activations with less interference from irrelevant regions. It consistently highlights the liver parenchyma in the Hepatorenal and Left Lobe planes, while producing more localized and coherent activation around the portal vein in the Transverse Plane. These results suggest that our method learns more anatomically grounded representations for ultrasound standard plane recognition.

5.3 Ablation Study

Ablation Study of Each Component. We conduct ablation studies on the LP2025 dataset, with the results reported in Table 3. Starting from the baseline, introducing Hierarchical Alignment leads to a clear performance improvement, demonstrating the benefit of hierarchical semantic supervision. Adding Coarse Alignment further improves the results, indicating that coarse-grained semantic information provides complementary supervision. Finally, incorporating Semantic Routing achieves the best performance, reaching 84.68% Accuracy and 79.36% F1-score. These results demonstrate the complementary contributions of the proposed components and the effectiveness of semantic-guided routing.

Table 3: Ablation Study on the LP2025 dataset.

Method	Acc(%)↑	F1 (%)↑
Baseline	80.73	73.03
+ Hierarchical Alignment	83.13	75.64
+ Coarse Alignment	83.65	77.24
+ Semantic Routing (Ours)	**84.68**	**79.36**

Ablation on Hierarchical Semantics. Table 4 evaluates the contribution of the three hierarchical semantic levels on the LP2025 dataset. Removing any individual semantic level consistently degrades performance, indicating that local anatomical, relational, and plane-specific semantics provide complementary information for standard plane recognition. Among them, removing the Plane-level semantics causes the largest performance degradation, with Accuracy and F1-score dropping to 83.23% and 75.53%, respectively, highlighting the importance of plane-specific semantic cues for fine-grained discrimination. The full configuration achieves the best performance, reaching 84.68% Accuracy and 79.36% F1-score, demonstrating the benefit of jointly modeling complementary semantic information across different levels.

Table 4: Ablation study of hierarchical semantic levels on the LP2025 dataset.

Semantic Configuration	Acc (%)↑	F1 (%)↑
w/o Local	84.38	78.67
w/o Relational	84.17	77.30
w/o Plane	83.23	75.53
Ours	**84.68**	**79.36**

t-SNE Visualization of Feature Discrimination. To further assess the discriminative quality of the learned representations, we visualize the feature embeddings using t-SNE. As shown in Figure 5, the baseline shows substantial inter-class overlap, whereas our method forms more compact intra-class clusters with clearer inter-class separation. This qualitative comparison suggests that the proposed semantic-guided hierarchical modeling encourages more structured feature representations and improves the discriminability of ultrasound standard plane recognition.

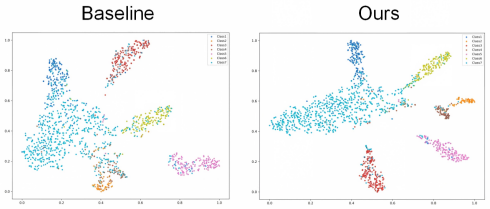


Figure 5: T-SNE visualization of feature distributions. (Left) Baseline model. (Right) Our method.

Ablation on Loss Weighting. To evaluate the effectiveness of the proposed adaptive loss weighting strategy, we compare three weighting schemes on the LP2025 dataset, as reported in Table 5. Equal weighting provides a strong baseline, while the best-performing fixed weighting scheme further improves the performance, suggesting that different semantic levels contribute unequally to standard plane recognition. The proposed adaptive weighting achieves the best overall performance, indicating that dynamically adjusting the contributions of different hierarchical alignment losses provides more effective semantic supervision than using a manually selected fixed weighting scheme.

Table 5: Effect of weighting strategies for hierarchical alignment losses on the LP2025 dataset.

Weighting Strategy	Acc(%)↑	F1(%)↑
Equal	84.06	78.19
Fixed	84.37	78.67
Adaptive (Ours)	**84.68**	**79.36**

6 Conclusion

In this paper, we propose **SHE-Net**, a semantic-guided hierarchical experts framework for ultrasound standard plane recognition. By introducing hierarchical vision-language alignment and semantic-aware expert routing, SHE-Net incorporates explicit anatomical semantics into visual representation learning and enables complementary modeling from local structures to global plane configurations. The hierarchical alignment strategy provides multi-level semantic supervision, while the expert routing mechanism adaptively promotes specialized feature learning for different anatomical patterns. We further construct and annotate a new kidney ultrasound standard plane dataset to facilitate evaluation under a different anatomical setting. Extensive experiments on three ultrasound standard plane datasets demonstrate consistent improvements over existing methods, validating the effectiveness and generalizability of the proposed framework. These results highlight the potential of combining anatomical semantics with adaptive expert specialization for learning more discriminative and structurally consistent representations, providing a promising direction for robust ultrasound standard plane recognition.

AI USE STATEMENT

Generative AI tools were used solely to assist with language polishing and improving the clarity and readability of the manuscript. No generative AI tools were used to generate experimental results, data, or scientific claims. All AI-assisted text was carefully reviewed and revised by the authors. The authors take full responsibility for the final content of this work.

ETHICS STATEMENT

This work uses anonymized ultrasound images for research on standard plane recognition. No personally identifiable information is included in the data used in this study. The proposed method is intended for research purposes and is not designed to replace professional clinical judgment.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the proposed method, training objectives, experimental settings, and evaluation protocols in the main paper and supplementary material. Additional implementation details and experimental results are provided to facilitate reproducibility. Code and relevant resources will be released upon acceptance, subject to applicable data and licensing restrictions.

REFERENCES

Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Perez-Garcia, Maximilian Ilse, Daniel C Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, et al. Learning to exploit temporal structure for biomedical vision-language processing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15016–15027, 2023.

Christian F Baumgartner, Konstantinos Kamnitsas, Jacqueline Matthew, Tara P Fletcher, Sandra Smith, Lisa M Koch, Bernhard Kainz, and Daniel Rueckert. Real-time detection and localisation of fetal standard scan planes in 2d freehand ultrasound. *arXiv preprint arXiv:1612.05601*, 2016.

Qing Cai, Fan Zhang, Cheng Zhang, Zhi Liu, et al. Semc: Structure-enhanced mixture-of-experts contrastive learning for ultrasound standard plane recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 2543–2551, 2026.

Jieneng Chen, Jieru Mei, Xianhang Li, Yongyi Lu, Qihang Yu, Qingyue Wei, Xiangde Luo, Yutong Xie, Ehsan Adeli, Yan Wang, et al. Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers. *Medical Image Analysis*, 97:103280, 2024.

Xiaohua Chen, Yucan Zhou, Dayan Wu, Chule Yang, Bo Li, Qinghua Hu, and Weiping Wang. Area: adaptive reweighting via effective area for long-tailed classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19277–19287, 2023.

Robert W Cootney. Ultrasound imaging: principles and applications in rodent research. *ILAR journal*, 42(3):233–247, 2001.

Mariachiara Di Cosmo, Maria Chiara Fiorentino, Francesca Pia Villani, Emanuele Frontoni, Gianluca Smerilli, Emilio Filippucci, and Sara Moccia. A deep learning approach to median nerve evaluation in ultrasound images of carpal tunnel inlet. *Medical & Biological Engineering & Computing*, 60(11):3255–3264, 2022.

Richard Droste, Yifan Cai, Harshita Sharma, Pierre Chatelain, Lior Drukker, Aris T Papageorghiou, and J Alison Noble. Ultrasound image representation learning by modeling sonographer visual attention. In *International conference on information processing in medical imaging*, pp. 592–604. Springer, 2019.

Aarthipoornima Elangovan and Thangaraja Jeyaseelan. Medical imaging modalities: a survey. In *2016 International Conference on emerging trends in engineering, technology and science (ICETETS)*, pp. 1–4. ieee, 2016.

Zhiwen Fan, Rishov Sarkar, Ziyu Jiang, Tianlong Chen, Kai Zou, Yu Cheng, Cong Hao, Zhangyang Wang, et al. M^3vit: Mixture-of-experts vision transformer for efficient multi-task learning with model-accelerator co-design. *Advances in Neural Information Processing Systems*, 35:28441–28457, 2022.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Juncheng Guo, Guanghua Tan, Fan Wu, Huaxuan Wen, and Kenli Li. Fetal ultrasound standard plane detection with coarse-to-fine multi-task learning. *IEEE Journal of Biomedical and Health Informatics*, 27(10):5023–5031, 2022.

Xumeng Han, Longhui Wei, Zhiyang Dou, Yingfei Sun, Zhenjun Han, and Qi Tian. Vimoe: An empirical study of designing vision mixture-of-experts. *IEEE Transactions on Image Processing*, 34:7209–7221, 2025.

Hussein Hazimeh, Zhe Zhao, Aakanksha Chowdhery, Maheswaran Sathiamoorthy, Yihua Chen, Rahul Mazumder, Lichan Hong, and Ed Chi. Dselect-k: Differentiable selection in the mixture of experts with applications to multi-task learning. *Advances in Neural Information Processing Systems*, 34:29335–29347, 2021.

Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3942–3951, 2021.

Jaehoon Jeong, Yi Hu, Soopil Kim, Jongseong Jang, Soonyoung Lee, and Sang Hyun Park. Ivaan: Instance-level vision-language alignment via attribute-guided text prompts generation for nuclei analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29305–29314, 2026.

Jiu Jiang, Qi Zhou, Haoying He, Junjian Zhang, and Chu He. Fine-grained vision-language alignment for medical image segmentation. *Available at SSRN 6188924*, 2026.

Jing Jiao, Jin Zhou, Xiaokang Li, Menghua Xia, Yi Huang, Lihong Huang, Na Wang, Xiaofan Zhang, Shichong Zhou, Yuanyuan Wang, et al. Usfm: A universal ultrasound foundation model generalized to tasks and organs towards label efficient image analysis. *Medical image analysis*, 96:103202, 2024.

Yan Jin, Mengke Li, Yang Lu, Yiu-ming Cheung, and Hanzi Wang. Long-tailed visual recognition via self-heterogeneous integration with knowledge excavation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 23695–23704, 2023.

Tsung-Wei Ke, Sangwoo Mo, and Stella Yu. Learning hierarchical image segmentation for recognition and by recognition. In *International Conference on Learning Representations*, volume 2024, pp. 39881–39919, 2024.

Frederick W Kremkau. *Sonography principles and instruments*. Elsevier Health Sciences, 2015.

Thunakala Bala Krishna and Priyanka Kokil. Standard fetal ultrasound plane classification based on stacked ensemble of deep learning models. *Expert Systems with Applications*, 238:122153, 2024.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*, 2020.

Qingqiu Li, Xiaohan Yan, Jilan Xu, Runtian Yuan, Yuejie Zhang, Rui Feng, Quanli Shen, Xiaobo Zhang, and Shujun Wang. Anatomical structure-guided medical vision-language pre-training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 80–90. Springer, 2024.

Chenyu Lian, Hong-Yu Zhou, Dongyun Liang, Jing Qin, and Liansheng Wang. Efficient medical vision-language alignment through adapting masked vision models. *IEEE Transactions on Medical Imaging*, 2025.

Yaru Liu, Jiangbo Pei, Zhu He, Guangjing Yang, Zhuqing Jiang, and Qicheng Lao. Medical language mixture of experts for improving medical image segmentation. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 2210–2216. IEEE, 2024.

Mohammad Ali Maraci, Raffaele Napolitano, Aris Papageorghiou, and J Alison Noble. Searching for structures of interest in an ultrasound video sequence. In *International Workshop on Machine Learning in Medical Imaging*, pp. 133–140. Springer, 2014.

Giovanna Migliorelli, Maria Chiara Fiorentino, Mariachiara Di Cosmo, Francesca Pia Villani, Adriano Mancini, and Sara Moccia. On the use of contrastive learning for standard-plane classification in fetal ultrasound imaging. *Computers in Biology and Medicine*, 174:108430, 2024.

Christopher L Moore and Joshua A Copel. Point-of-care ultrasonography. *New England Journal of Medicine*, 364(8):749–757, 2011.

Srinivasa Rao Nandam, Sara Atito, Zhenhua Feng, Josef Kittler, and Muhammad Awais. Text augmented correlation transformer for few-shot classification & segmentation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25357–25366. IEEE, 2025.

Bharath Srinivas Prabakaran, Paul Hamelmann, Erik Ostrowski, and Muhammad Shafique. 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.

Bin Pu, Kenli Li, Shengli Li, and Ningbo Zhu. Automatic fetal ultrasound standard plane recognition based on deep learning and iiot. *IEEE Transactions on Industrial Informatics*, 17(11):7771–7780, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems*, 34:8583–8595, 2021.

Laurent Julien Salomon, Zarko Alfirevic, Vincenzo Berghella, C Bilardo, Edgar Hernandez-Andrade, Synnøve Lian Johnsen, K Kalache, K-Y Leung, G Malinger, H Munoz, et al. Practice guidelines for performance of the routine mid-trimester fetal ultrasound scan. *Ultrasound in Obstetrics & Gynecology*, 37(1), 2011.

Jo Schlemper, Ozan Oktay, Michiel Schaap, Mattias Heinrich, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention gated networks: Learning to leverage salient regions in medical images. *Medical image analysis*, 53:197–207, 2019.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

Jacqueline K Spencer and Ronald S Adler. Utility of portable ultrasound in a community in ghana. *Journal of Ultrasound in Medicine*, 27(12):1735–1743, 2008.

Sara Tehsin, Hend Alshaya, Wided Bouchelligua, and Inzamam Mashood Nasir. Hybrid state–space and vision transformer framework for fetal ultrasound plane classification in prenatal diagnostics. *Diagnostics*, 15(22):2879, 2025.

Yipei Wang, Qianye Yang, Lior Drukker, Aris Papageorghiou, Yipeng Hu, and J Alison Noble. Task model-specific operator skill assessment in routine fetal ultrasound scanning. *International Journal of Computer Assisted Radiology and Surgery*, 17(8):1437–1444, 2022a.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3876–3887, 2022b.

HN Xie, Nan Wang, Miao He, LH Zhang, HM Cai, JB Xian, MF Lin, Jv Zheng, and YZ Yang. Using deep-learning algorithms to classify fetal brain ultrasound images as normal or abnormal. *Ultrasound in Obstetrics & Gynecology*, 56(4):579–587, 2020.

Yijun Yang, Huazhu Fu, Angelica I Aviles-Rivero, Carola-Bibiane Schönlieb, and Lei Zhu. Diffmic: Dual-guidance diffusion network for medical image classification. In *International conference on medical image computing and computer-assisted intervention*, pp. 95–105. Springer, 2023.

Mohammad Yaqub, Brenda Kelly, Aris T Papageorghiou, and J Alison Noble. A deep learning solution for automatic fetal neurosonographic diagnostic plane verification using clinical standard constraints. *Ultrasound in medicine & biology*, 43(12):2925–2933, 2017.

Tianxiang Yu, Po-Hsiang Tsui, Denis Leonov, Shuicai Wu, Guangyu Bin, and Zhuhuang Zhou. Lpc-sononet: A lightweight network based on sononet and light pyramid convolution for fetal ultrasound standard plane detection. *Sensors*, 24(23):7510, 2024.

Zhen Yu, Dong Ni, Siping Chen, Shengli Li, Tianfu Wang, and Baiying Lei. Fetal facial standard plane recognition via very deep convolutional networks. In *2016 38th annual international conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 627–630. IEEE, 2016.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine learning for healthcare conference*, pp. 2–25. PMLR, 2022.

Lei Zhao, Kenli Li, Bin Pu, Jianguo Chen, Shengli Li, and Xiangke Liao. An ultrasound standard plane detection model of fetal head based on multi-task learning and hybrid knowledge graph. *Future Generation Computer Systems*, 135:234–243, 2022.

A Appendix

This supplementary material provides additional details and experimental results that complement the main manuscript of **SHE-Net**. The content is organized as follows:

- **Dataset Details:** We provide detailed statistics and descriptions of the FPUS23 and LP2025 datasets, with particular emphasis on intra-class variation and inter-class similarity. We further describe the construction of hierarchical semantic text embeddings under clinical guidance, including dataset-specific coarse-grained descriptions and plane-level fine-grained semantics, together with representative examples.
- **Additional Experimental Results:** We present additional experimental results to further evaluate the effectiveness of the proposed method.

A.1 Dataset Details

A.1.1 FPUS23 Dataset

Plane	**Abbreviation**	**Number**
Abdomen	AB	711
Brain	BR	3092
Femur	FE	1040
Maternal	MA	1626
Thorax	TH	1718
Non-Standard Plane	NSP	4213
Total	–	12400

Table 6: Distribution of images in the FPUS23 dataset by standard plane.

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 7: Distribution of images in the LP2025 dataset across six liver standard planes and the Non-Standard Plane (NSP) category.

FPUS23 is a fetal ultrasound benchmark for standard plane recognition, covering several representative views used in prenatal ultrasound examination, including abdomen, brain, femur, maternal, and thorax planes, together with a Non-Standard Plane (NSP) category. The detailed sample distribution is reported in Table 6. The dataset exhibits substantial visual variability within the same plane due to differences in gestational stage, fetal motion, scanning conditions, and anatomical appearance. Meanwhile, different planes may share similar anatomical structures or visual patterns, resulting in considerable inter-class similarity. These characteristics make FPUS23 a suitable benchmark for evaluating the ability of standard plane recognition models to capture discriminative anatomical and contextual cues.

A.1.2 LP2025 Dataset

LP2025 is a publicly available dataset for liver standard plane recognition in abdominal ultrasound. It contains six anatomically defined standard planes, namely FHP1, FHP2, LLP, RLP, LPV-S, and HRP, together with a Non-Standard Plane (NSP) category, as summarized in Table 7.

Compared with fetal ultrasound, liver standard plane recognition places greater emphasis on fine-grained anatomical discrimination, where subtle differences in the visibility and spatial arrangement of anatomical structures can determine the target plane. The dataset is annotated by experienced sonographers following routine clinical practice. Variations in scanning angle, anatomical morphology, and image quality further increase the visual diversity within each category, while the anatomical similarities between neighboring planes lead to substantial inter-class similarity.

A.1.3 Construction of Multi-level Semantic Text Embeddings

To facilitate fine-grained semantic alignment, we construct structured textual embeddings at multiple semantic levels under clinical guidance. Specifically, our semantic design follows a coarse-to-fine hierarchy, consisting of dataset-specific coarse-grained descriptions and fine-grained descriptions at the local, relational, and plane levels, which provide complementary semantic cues for representation learning. At the coarse-grained level, we construct dataset-specific semantic descriptors that characterize the overall anatomical context and examination focus. For each dataset, namely FPUS23, LP2025, and KU2026, a dedicated set of coarse-grained textual descriptions is developed to reflect its specific imaging domain, such as fetal anatomy, liver structures, and renal regions. These descriptions provide global semantic priors that help distinguish different anatomical systems and imaging contexts.

At the fine-grained level, we construct detailed semantic descriptions for each standard plane to capture subtle and discriminative anatomical characteristics. Specifically, the descriptions cover three complementary aspects: *local anatomical structures*, such as vessels and parenchyma; *anatomical relationships*, such as the relative spatial arrangement between structures; and *plane-specific characteristics*, such as typical appearance patterns and key anatomical landmarks. For each dataset, a dedicated collection of fine-grained descriptions is constructed, resulting in three dataset-specific semantic groups corresponding to FPUS23, LP2025, and KU2026, respectively. All semantic descriptions are developed in collaboration with experienced clinicians to ensure anatomical accuracy and clinical relevance. The resulting multi-level textual embeddings form a structured semantic



Figure 6: Representative examples of fine-grained semantic text descriptions constructed for the KU2026 kidney ultrasound dataset. Similar hierarchical descriptions are constructed for FPUS23 and LP2025.

space, in which coarse-grained descriptions provide global contextual guidance, while fine-grained descriptions support precise anatomical discrimination. Figure 6 presents representative examples of the fine-grained descriptions constructed for the KU2026 dataset, illustrating their hierarchical organization and anatomical interpretability.

A.2 Additional Experimental Results

A.2.1 Additional Qualitative Comparison

Figures 7 and 8 present additional Grad-CAM visualizations on the FPUS23 and KU2026 datasets, respectively. Across different fetal and kidney ultrasound standard planes, our method generally produces more localized and anatomically coherent activations, with less interference from irrelevant background regions. Compared with the baseline methods, it more consistently highlights task-relevant anatomical structures while suppressing non-discriminative regions. These qualitative results further demonstrate the ability of the proposed semantic-guided framework to learn anatomically grounded visual representations across different ultrasound imaging domains.

A.2.2 Additional Ablation Study

Ablation on Coarse Weighting. Table 8 investigates the effect of the coarse-level alignment weight λ_{coarse}. Among the evaluated settings, $\lambda_{\text{coarse}} = 0.5$ achieves the best performance, with 84.68% Accuracy and 79.36% F1-score. Both smaller weights of 0.1 and 0.2 and the larger weight of 1.0 result in lower performance. These results suggest that an appropriate balance between coarse-level

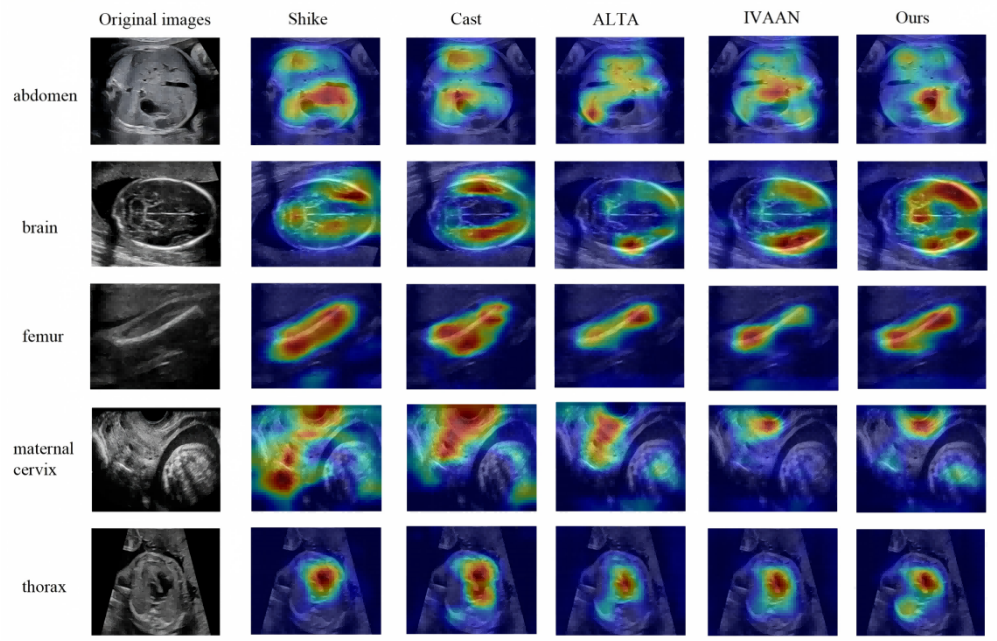


Figure 7: Grad-CAM visualizations of our method and representative baselines on the FPUS23 dataset. The results are shown for representative fetal ultrasound standard planes, illustrating the differences in spatial attention and anatomical localization across methods.

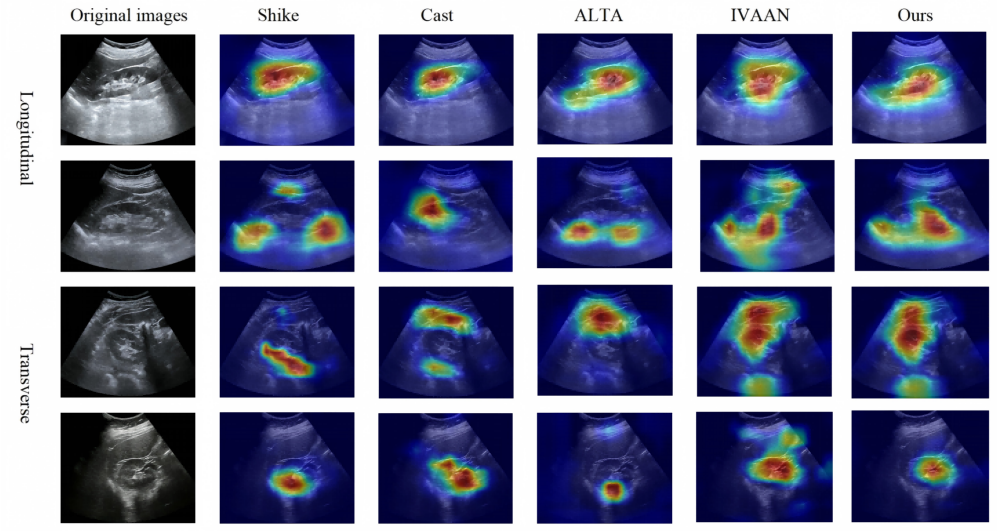


Figure 8: Grad-CAM visualizations of our method and representative baselines on the KU2026 dataset. The visualizations demonstrate the anatomical regions emphasized by different methods for representative kidney ultrasound standard planes.

and fine-grained semantic supervision is important for effective representation learning. Therefore, we set $\lambda_{\text{coarse}} = 0.5$ in all experiments.

Ablation on Route Weighting. Table 9 examines the effect of the routing constraint weight λ_{route}. Among the evaluated settings, $\lambda_{\text{route}} = 0.2$ achieves the best performance, reaching 84.68% Accuracy and 79.36% F1-score. Increasing the weight to 0.5 or 1.0 leads to lower performance, with a more pronounced decrease in F1-score. These results suggest that a moderate routing constraint

Table 8: Ablation study of the coarse-level alignment weight λ_{coarse} on the LP2025 dataset.

λ_{coarse}	Accuracy (%)↑	F1-score (%)↑
0.1	84.19	77.27
0.2	84.23	76.72
0.5	**84.68**	**79.36**
1.0	84.33	78.17

Table 9: Ablation study of the routing constraint weight λ_{route} on the LP2025 dataset.

λ_{route}	Accuracy (%)↑	F1-score (%)↑
0.1	84.32	77.45
0.2	**84.68**	**79.36**
0.5	84.44	76.60
1.0	84.21	77.25

provides a better balance between routing regularization and the primary recognition objective. Accordingly, we set $\lambda_{\text{route}} = 0.2$ in all experiments.