HAR-MoE: Hierarchical Affinity-aware Routing Mixture-of-Experts for Unified Multi-Organ Ultrasound Analysis

Yajun Mu[1,2], Yong Chen[2], Shuhan Yang[3], Jun Gao[2], Qicheng Lao[4], Kang Li[1,2*], and Qingbo Kang[2 *]

[1] School of Computing and Artificial Intelligence, Southwest Jiaotong University
[2] West China Biomedical Big Data Center, West China Hospital, Sichuan University
[3] Pittsburgh Institute, Sichuan University, Chengdu, China
[4] School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China

Abstract. Developing unified models for multi-organ ultrasound analysis is inherently challenging due to the conflicting optimization dynamics between dense pixel-level segmentation and sparse image-level classification. Existing Mixture-of-Experts (MoE) architectures often fail in this heterogeneous setting, suffering from *expert dilution*, where purely data-driven routers are dominated by dense segmentation gradients, leading to the collapse of experts intended for sparse tasks. To address this, we propose HAR-MoE (**H**ierarchical **A**ffinity-aware **R**outing MoE). Unlike monolithic routers, HAR-MoE introduces a two-stage hierarchy: a task-conditioned coarse stage first screens experts to balance gradient density, followed by an organ-conditioned fine stage for anatomical specialization. We further incorporate learnable affinity priors to explicitly decouple optimization logic from representation learning. Extensive experiments on 15 heterogeneous ultrasound datasets spanning 7 organs demonstrate that HAR-MoE consistently achieves superior performance across both segmentation and classification tasks. Notably, our framework boosts fine-grained diagnosis by over 16% (F1-score), confirming its efficacy in mitigating negative transfer in unified multi-organ ultrasound analysis. Code: https://github.com/kzxl0721/HAR-MoE.

Keywords: Ultrasound Image Analysis · Mixture-of-Experts · Multi-Task Learning · Unified Model.

1 Introduction

Ultrasound imaging is vital for clinical diagnosis due to its real-time, safe, and cost-effective nature [16,25]. There is growing interest in *unified models* for multi-organ analysis, performing tasks from dense pixel-level segmentation to sparse image-level classification [11]. However, these tasks differ in optimization dynamics: segmentation requires dense pixel-wise gradients, while classification relies on

* Corresponding authors: likang@wchscu.cn, kangqingbokb@gmail.com.

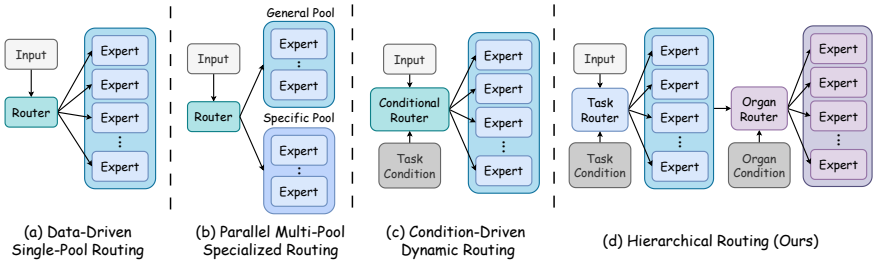

Fig. 1. Comparison of MoE routing paradigms. Conventional (a) single-pool, (b) parallel-pool, and (c) condition-driven routing strategies rely on flat decisions. In contrast, our (d) hierarchical routing employs two-stage task and organ condition for structured, precise expert specialization.

sparse semantic reasoning [22,30]. Conflating these heterogeneous tasks within a unified model often leads to negative transfer, where dominant dense gradients overwhelm sparse ones, causing mode collapse in classification.

Mixture-of-Experts (MoE) enables conditional computation [24], yet current designs struggle with gradient conflicts in medical imaging. Data-driven single-pool routing (Fig. 1a) suffers from *expert dilution*, where dense segmentation gradients dominate routing, leaving classification experts under-trained [30]. While parallel multi-pool (Fig. 1b) or condition-driven strategies (Fig. 1c) alleviate interference, they often hinder shared representation learning or entangle optimization logic with anatomical appearance [29,13]. Crucially, these limitations stem from a **structural misalignment**: task identity governs *optimization dynamics*, whereas organ identity reflects *anatomical patterns*. Conflating these distinct inductive roles within flat routing spaces exacerbates conflict, necessitating explicit decoupling of task objectives from anatomical variation (Fig. 1d).

To this end, we propose HAR-MoE: **H**ierarchical **A**ffinity-aware **R**outing Mixture-of-Experts, a unified framework that **decouples** task optimization from anatomical representation. Instead of a monolithic routing process, HAR-MoE employs a two-stage hierarchy: a task-conditioned coarse stage first filters experts to mitigate gradient dominance, followed by an organ-conditioned fine stage for anatomical specialization. We further introduce learnable affinity priors to stabilize routing, ensuring sparse classification receives dedicated expert capacity. Our main contributions are:

1. We identify *expert dilution* driven by gradient asymmetry as the primary cause of negative transfer in unified ultrasound MoEs. Our analysis reveals that simply scaling up expert capacity fails to resolve this conflict, necessitating a structural decoupling of optimization objectives.
2. We propose a hierarchical affinity-aware routing mechanism that explicitly disentangles task optimization logic from anatomical representation. By incorporating learnable affinity priors, our framework ensures stable, semantically consistent expert specialization without the need for explicit routing supervision.

3. Extensive experiments on 15 heterogeneous datasets demonstrate that HAR-MoE achieves state-of-the-art performance across all tasks. Crucially, it resolves expert dilution, boosting fine-grained breast subtype classification by over 16% in F1-score while maintaining superior accuracy in dense segmentation.

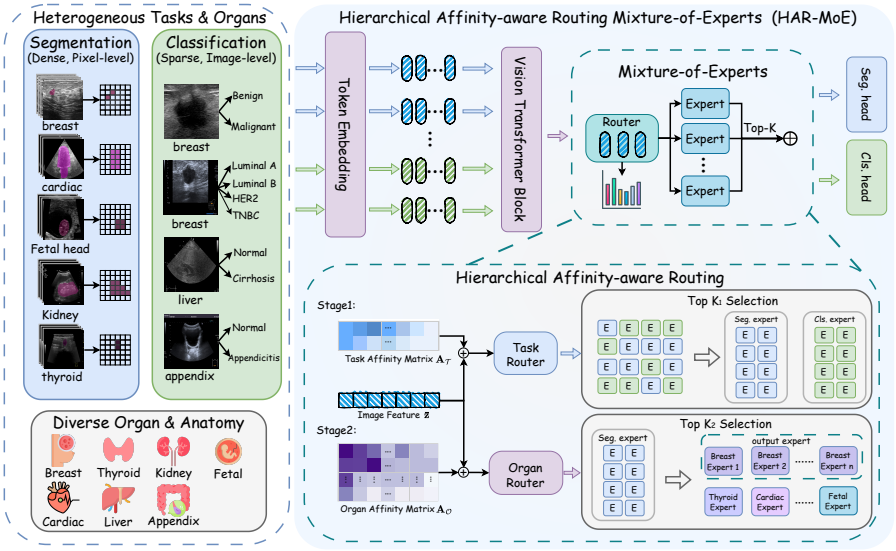

Fig. 2. Overview of the proposed HAR-MoE. It unifies heterogeneous dense segmentation and sparse classification across diverse organs within a shared ViT. To mitigate gradient conflicts, standard FFNs are replaced by a two-stage hierarchical routing MoE: a task-conditioned coarse stage screens candidate experts for optimization decoupling, followed by an organ-conditioned fine stage for precise anatomical specialization.

2 Methods

2.1 Problem Definition and Overall Framework

We formulate unified multi-organ ultrasound analysis under heterogeneous supervision, comprising dense pixel-level segmentation ($\mathcal{T}_{\text{seg}}$) and sparse image-level classification ($\mathcal{T}_{\text{cls}}$). Our objective is a shared Vision Transformer (ViT) backbone (Fig. 2) that extracts task-agnostic representations while conditioning computation on *optimization objectives* and *anatomical contexts*. To mitigate negative transfer from conflicting gradient dynamics, we replace the Feed-Forward Networks (FFNs) in selected blocks with HAR-MoE layers. Unlike monolithic routing, HAR-MoE directs tokens through an expert hierarchy guided by learned affinity priors, followed by lightweight task-specific prediction heads.

2.2 Hierarchical Affinity-Aware Routing

The core of HAR-MoE is a routing mechanism that explicitly decomposes expert selection into task-level and organ-level decisions. By disentangling these orthogonal factors via a two-stage strategy, we prevent dense segmentation gradients from overwhelming sparse classification tasks.

Factorized Dynamic Affinity Modeling Instead of relying solely on a purely data-driven gating network, we inject structured inductive biases into the routing mechanism through decoupled affinity modeling. Let T and C denote the number of task and organ categories, respectively. Let $t \in \{1, \ldots, T\}$ and $c \in \{1, \ldots, C\}$ denote the corresponding task and organ indices. Let M denote the total number of experts, with $m \in \{1, \ldots, M\}$ indexing individual experts. We introduce two learnable affinity matrices:

- **Task Affinity Matrix** $\mathbf{A}_{\mathcal{T}} \in \mathbb{R}^{T \times M}$, where $\mathbf{A}_{\mathcal{T}}[t, m]$ models the global compatibility between task t and expert m.
- **Organ Affinity Matrix** $\mathbf{A}_{\mathcal{O}} \in \mathbb{R}^{C \times M}$, where $\mathbf{A}_{\mathcal{O}}[c, m]$ captures the preference of expert m for organ c.

From a probabilistic viewpoint, expert routing can be interpreted as approximating the posterior distribution $p(m \mid \mathbf{z}, t, c)$ over the discrete expert variable m, conditioned on image features $\mathbf{z}$, task identity t, and organ identity c. The routing logits adopt a log-additive form:

$$\log p(m \mid \mathbf{z}, t, c) \propto f(\mathbf{z})_m + \mathbf{A}_{\mathcal{T}}[t, m] + \mathbf{A}_{\mathcal{O}}[c, m], \tag{1}$$

where $f(\mathbf{z})_m$ denotes the instance-dependent score produced by the router MLP. Under a factorized prior assumption $p(m \mid t, c) \approx p(m \mid t)\, p(m \mid c)$, the affinity matrices can be interpreted as log-priors:

$$\mathbf{A}_{\mathcal{T}}[t, m] \approx \log p(m \mid t), \quad \mathbf{A}_{\mathcal{O}}[c, m] \approx \log p(m \mid c). \tag{2}$$

This structured decomposition decouples optimization-related preferences (task) from anatomical specialization (organ), providing a stable inductive bias during early training.

Two-Stage Hierarchical Routing Let $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_L] \in \mathbb{R}^{L \times D}$ denote the token embeddings produced by the transformer block. We first aggregate global contextual information via average pooling to obtain an image-level representation $\bar{\mathbf{z}} = \frac{1}{L} \sum_{i=1}^{L} \mathbf{z}_i$. Expert routing is performed at the image level and shared across all tokens, since both task identity and organ identity are image-level conditions in our setting.

Stage 1: Task-Conditioned Coarse Routing. This stage aims to screen experts based on optimization objectives to mitigate gradient conflict. For a sample associated with task index t, we compute the coarse routing logits $\mathbf{l}_{\text{task}} \in \mathbb{R}^M$ and select the top-K_1 experts to form a candidate set $\mathcal{S}_{\text{task}}$ as follows:

$$\mathbf{l}_{\text{task}} = f_\phi(\bar{\mathbf{z}}) + \mathbf{A}_{\mathcal{T}}[t], \quad \mathcal{S}_{\text{task}} = \text{TopK}(\mathbf{l}_{\text{task}}, K_1). \tag{3}$$

Stage 2: Organ-Conditioned Fine Routing. This stage refines the selection based on anatomical characteristics. Using the organ index c, we compute fine-grained logits $\mathbf{l}_{\text{organ}} \in \mathbb{R}^M$:

$$\mathbf{l}_{\text{organ}} = f_\psi(\bar{\mathbf{z}}) + \mathbf{A}_{\mathcal{O}}[c], \tag{4}$$

where $f_\phi(\cdot)$ and $f_\psi(\cdot)$ are lightweight MLP routers.

To ensure hierarchical consistency, we mask experts strictly outside the candidate set $\mathcal{S}_{\text{task}}$ (by setting their logits to $-\infty$). The gating weights are then computed via a temperature-scaled Softmax over the valid candidates:

$$g_m = \frac{\exp(\mathbf{l}_{\text{organ},m}/\tau)}{\sum_{j\in\mathcal{S}_{\text{task}}} \exp(\mathbf{l}_{\text{organ},j}/\tau)}, \quad \forall m \in \mathcal{S}_{\text{task}}, \tag{5}$$

where $\tau = 1$ is a temperature hyper-parameter. Finally, we select the top-K_2 experts from this subset to form the final active set $\mathcal{S}_{\text{final}}$, which reduces unconstrained expert switching compared with hard Top-K over the full expert pool. . The MoE output is computed via residual aggregation:

$$\mathbf{Y} = \mathbf{Z} + \sum_{m\in\mathcal{S}_{\text{final}}} g_m \, E_m(\mathbf{Z}), \tag{6}$$

where $E_m(\cdot)$ denotes the m-th SwiGLU-based expert network [23].

2.3 Task-Conditional Training Objective

Given the heterogeneous nature of the tasks, HAR-MoE adopts a task-conditional alternating optimization strategy. In each iteration, a mini-batch is sampled from a specific task dataset, and the loss function is dynamically adjusted:

$$\mathcal{L}_{\text{total}} = \begin{cases} \mathcal{L}_{\text{seg}} + \lambda\mathcal{L}_{\text{aux}}, & \text{if } \mathcal{T}_{\text{seg}}, \\ \mathcal{L}_{\text{cls}} + \lambda\mathcal{L}_{\text{aux}}, & \text{if } \mathcal{T}_{\text{cls}}. \end{cases} \tag{7}$$

Here, $\mathcal{L}_{\text{seg}}$ is the combination of Dice and Cross-Entropy loss for segmentation, $\mathcal{L}_{\text{cls}}$ is the standard Cross-Entropy loss for classification, and $\mathcal{L}_{\text{aux}}$ denotes the MoE load-balancing regularization term weighted by λ. This formulation avoids manual weighting between dense and sparse losses, further mitigating negative transfer.

3 Experiments

3.1 Experimental Setup

Datasets and Tasks We evaluate HAR-MoE on the UUSIC25 benchmark [14], a large-scale heterogeneous ultrasound suite from the UUSIC 2025 Challenge [5]. As summarized in Table 1, it spans 7 organs and covers two task types: image-level classification and pixel-level segmentation. Since the official test set is withheld, we adopt a reproducible re-split protocol, using the original validation set as test data and further dividing the training set into training and validation subsets, resulting in a consistent 7:1:2 split for all experiments.

[5] https://uusic2025.github.io

Table 1. Data summary.

Task	Organ	Target	Num.	Ref.
Cls.	Breast	Malignancy	2,672	[1,6]
	Breast	Subtype (4-class)	236	[14]
	Appendix	Appendicitis	1,042	[19]
	Liver	Steatosis	550	[3]
Seg.	Breast	Lesion	3,470	[1,6,31,14]
	Cardiac	Ventricle	576	[12,14]
	Fetal	Head	1,120	[8,14]
	Kidney	Kidney	554	[26,14]
	Thyroid	Nodule	894	[20,14]

Table 2. Impl. settings.

Setting	Value
Backbone	ViT-B [2]
Input size	224×224
MoE layers	Blocks 6, 9, 12
Experts	M=16
Top-K	K_1=8, K_2=4
Optimizer	AdamW [18]
LR	1×10^{-4}
Batch size	32
Epochs	100
λ (Eq. 7)	0.01

Table 3. Quantitative comparison across segmentation and classification tasks.

Method	Venue	Seg.		Cls. (Binary)			Cls. (Breast Subtype)		
		Dice ↑	mIoU ↑	ACC ↑	F1 ↑	AUC ↑	ACC ↑	F1 ↑	AUC ↑
U-Net [21]	MICCAI'15	83.08	72.41	–	–	–	–	–	–
nnUNet [9]	Nat.Methods'21	84.67	74.52	–	–	–	–	–	–
TransUNet [4]	MIA'24	78.54	66.03	–	–	–	–	–	–
M^{4}oE-Seg. [10]	MICCAI'24	74.54	63.25	–	–	–	–	–	–
ResNet-101 [7]	CVPR'16	–	–	75.39	62.43	72.88	66.67	19.62	60.19
Swin-T [17]	ICCV'21	–	–	75.29	68.29	75.63	64.58	19.62	57.42
ViT-B [2]	ICLR'21	–	–	83.46	80.45	87.64	62.75	19.28	53.34
UniUSNet [13]	BIBM'24	86.15	77.10	88.15	82.76	92.73	52.08	27.16	45.38
MTANet [15]	TMI'24	81.94	73.63	87.75	79.33	87.27	50.00	41.66	65.82
M^{4}oE-MTL [27]	ICLR'25	82.65	70.85	88.85	85.50	90.08	63.64	19.44	69.14
TaskExpert [28]	ICCV'23	83.56	73.05	86.64	81.32	87.87	62.50	37.45	68.06
AdaMV-MoE [5]	ICCV'23	87.90	79.45	91.20	87.56	90.92	68.75	37.06	68.24
ViT-MTL (Baseline)	–	83.99	74.24	79.54	74.04	80.62	64.58	19.62	57.22
HAR-MoE (Ours)	–	**88.33**	**80.01**	**92.59**	**90.05**	**95.09**	**76.47**	**58.12**	**71.29**

Implementation Settings HAR-MoE is implemented in PyTorch and trained on NVIDIA V100 GPUs. Hierarchical affinity-aware MoE layers are inserted into selected transformer blocks of the ViT backbone. Detailed hyperparameters and training configurations are summarized in Table 2. Single-task baselines are trained independently on their corresponding datasets, whereas all unified methods are jointly trained across all datasets and evaluated on the same held-out splits.

3.2 Quantitative Results

Table 3 compares HAR-MoE with representative single-task and unified multi-task learning (MTL) models, including a ViT-MTL baseline built on the same architecture. We evaluate all methods on segmentation, binary classification, and challenging fine-grained breast subtype classification.

Overall, HAR-MoE consistently improves performance across all tasks, demonstrating strong generalization under heterogeneous supervision. In segmentation,

Table 4. Ablation of routing strategies and affinity priors. Unlike *Flat Routing + Affinity* which applies priors globally, our hierarchical method disentangles them into task and organ stages.

Model Variant	Affinity Priors		Seg.		Cls. (Binary)			Cls. (Subtype)		
	Task	Organ	Dice ↑	IoU ↑	Acc ↑	F1 ↑	AUC ↑	Acc ↑	F1 ↑	AUC ↑
Baselines & Flat Routing										
ViT-MTL (No MoE)	–	–	83.99	74.24	79.54	74.04	80.62	64.58	19.62	57.22
Flat Routing	–	–	85.21	76.17	86.45	84.53	84.84	61.54	19.05	68.87
Flat Routing + Affinity	✓	✓	87.24	79.42	91.82	88.58	94.28	61.54	32.17	59.91
Hierarchical Routing (Ours)										
Hierarchical Routing	–	–	86.94	77.84	88.29	86.34	91.35	64.58	35.95	69.87
+ Task Affinity	✓	–	87.38	78.48	90.57	86.20	92.97	62.74	49.37	65.68
+ Organ Affinity	–	✓	87.35	77.18	89.46	86.29	92.38	63.42	51.73	67.92
HAR-MoE (Full)	✓	✓	**88.33**	**80.01**	**92.59**	**90.05**	**95.09**	**76.47**	**58.12**	**71.29**

Table 5. Ablation of expert configuration M (K_1, K_2: active experts per stage) and the MoE load-balancing regularization weight λ (Eq. 7).

$M(K_1, K_2)$	Params (M) Total	Active	FLOPs (G)	Seg. Dice	Bin. AUC	Sub. F1
64 (32, 16)	1511	415	81.0	87.78	**95.41**	50.63
32 (16, 8)	831	245	47.5	87.96	94.41	54.03
16 (8, 4)	491	159	30.8	**88.33**	95.09	**58.12**
8 (4, 2)	321	117	22.4	88.05	94.40	56.98

λ	Seg. Dice	Bin. AUC	Sub. F1
0.1	86.06	92.03	54.63
0.01	**88.33**	**95.09**	**58.12**
0.001	87.64	92.61	53.62
1e-4	87.85	93.28	56.37

HAR-MoE attains a Dice score of 88.33%, outperforming AdaMV-MoE (87.90%) and UniUSNet (86.15%). This indicates precise anatomical delineation is maintained despite concurrent classification optimization. Similarly, for binary classification, HAR-MoE shows superior discriminative capability with 95.09% AUC, surpassing second-best UniUSNet by 2.36%.

The most substantial gains occur in breast subtype classification, where standard unified models typically fail. As shown in Table 3, baselines like ViT-MTL, M^{4}oE-MTL suffer from mode collapse (F1 ≈19%) due to dominant segmentation gradients. While recent MoE methods (TaskExpert, AdaMV-MoE) struggle to exceed 40% F1, HAR-MoE achieves 58.12%, outperforming the second-best MTANet by 16.46%. These results validate that hierarchical affinity-aware routing effectively decouples optimization logic, ensuring dedicated capacity for sparse tasks to mitigate negative transfer.

3.3 Ablation Studies

Impact of Hierarchical Routing and Affinity Priors Table 4 shows that hierarchical routing consistently outperforms flat routing, especially in fine-grained classification. Compared to Flat Routing, our hierarchical routing improves breast subtype F1 by 16.9%, indicating that decoupling task- and organ-level routing effectively mitigates optimization conflict under heterogeneous su-

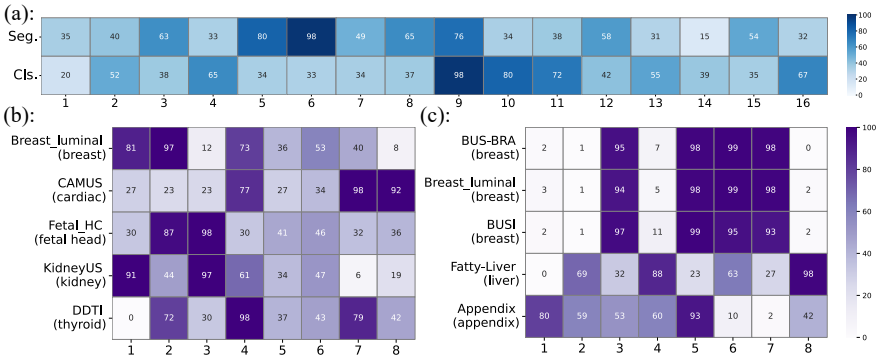

Fig. 3. Hierarchical expert activation in HAR-MoE: (a) task-level routing; (b) organ-level routing for segmentation; and (c) organ-level routing for classification.

pervision. While affinity priors improve flat routing, they remain inferior to hierarchical routing. Within our framework, task affinity stabilizes optimization objectives, while organ affinity enhances fine-grained discrimination. Combining both yields the best overall performance, demonstrating their synergy in structured expert specialization.

Impact of Expert Configuration and Efficiency Table 5 evaluates expert capacity and load-balancing regularization (Eq. 7). A configuration with $M = 16$ experts balances performance and efficiency. Notably, increasing capacity to $M = 64$ degrades breast subtype performance, providing empirical evidence for expert dilution under heterogeneous supervision. The framework is robust to the load-balancing weight λ, with $\lambda = 0.01$ yielding stable performance across tasks and preventing expert collapse.

3.4 Visualization

Fig. 3 visualizes the hierarchical expert activation. At the task level (Fig. 3(a)), segmentation and classification activate largely disjoint experts, indicating effective task-level disentanglement. At the organ level (Fig. 3(b,c)), distinct experts are selected for different anatomy. Notably, the three breast datasets exhibit highly consistent activation (Fig. 3(c)), suggesting routing is driven by shared organ semantics rather than dataset biases. These structured patterns confirm effective task-organ decoupling.

4 Conclusion

In this paper, we identify that the primary bottleneck in unified ultrasound analysis lies in the gradient asymmetry between dense and sparse tasks, which leads

to expert dilution in conventional MoE models. We introduce HAR-MoE, which resolves this through a hierarchical routing strategy and learnable affinity priors. Our framework effectively disentangles optimization-specific logic from anatomical representation, ensuring that fine-grained diagnostic tasks are not overwhelmed by dominant segmentation gradients. The significant performance gains across 15 datasets validate the scalability and robustness of our approach. We believe the principle of hierarchical decoupling serves as a promising paradigm for broader multi-task medical AI systems handling heterogeneous supervision. Future work may explore smoother routing or hybrid token-image routing to further improve stability for highly heterogeneous lesions.

Acknowledgment. This study was supported by Noncommunicable Chronic Diseases–National Science and Technology Major Project (2026ZD0555600 and 2026ZD0555601) and Chengdu Science and Technology Program (2026-XT00-00016-GX)

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

References

1. Al-Dhabyani, W., et al.: Dataset of breast ultrasound images. Data in Brief **28**, 104863 (2020)
2. Alexey Dosovitskiy, Lucas Beyer, A.K., et al.: An image is worth 16x16 words: Transformers for image recognition at scale (2021), https://arxiv.org/abs/2010.11929
3. Byra, M., et al.: Transfer learning with deep convolutional neural network for liver steatosis assessment in ultrasound images. International Journal of Computer Assisted Radiology and Surgery **13**(12), 1895–1903 (2018)
4. Chen, J., Mei, J., Li, X., et al.: Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers. Medical Image Analysis **97**, 103280 (2024)
5. Chen, T., et al.: Adamv-moe: Adaptive multi-task vision mixture-of-experts. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 17300–17311 (2023). https://doi.org/10.1109/ICCV51070.2023.01591
6. Gómez-Flores, W., et al.: Bus-bra: A breast ultrasound dataset for assessing computer-aided diagnosis systems. Medical Physics **51**(4), 3110–3123 (2024)
7. He, K., Zhang, X., Ren, S., et al.: Deep residual learning for image recognition (2015), https://arxiv.org/abs/1512.03385
8. van den Heuvel, T.L., et al.: Automated measurement of fetal head circumference using 2d ultrasound images. PLoS ONE **13**(8), e0200412 (2018)
9. Isensee, F., et al.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods **18**(2), 203–211 (2021)
10. Jiang, Y., Shen, Y.: M4oE: A Foundation Model for Medical Multimodal Image Segmentation with Mixture of Experts . In: proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. vol. LNCS 15012. Springer Nature Switzerland (October 2024)

11. Kang, Q., et al.: Thyroid nodule segmentation and classification in ultrasound images through intra-and inter-task consistent learning. Medical image analysis **79**, 102443 (2022)
12. Leclerc, S., et al.: Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. IEEE Transactions on Medical Imaging **38**(9), 2198–2210 (2019)
13. Lin, Z., et al.: Uniusnet: A promptable framework for universal ultrasound disease prediction and tissue segmentation (2024), https://arxiv.org/abs/2406.01154
14. Lin, Z., et al.: Diagnostic performance of universal-learning ultrasound ai across multiple organs and tasks: the uusic25 challenge. arXiv preprint arXiv:2512.17279 (2025)
15. Ling, Y., et al.: Mtanet: Multi-task attention network for automatic medical image segmentation and classification. IEEE Transactions on Medical Imaging **43**(2), 674–685 (2024). https://doi.org/10.1109/TMI.2023.3317088
16. Litjens, G., et al.: A survey on deep learning in medical image analysis. Medical Image Analysis **42**, 60–88 (2017)
17. Liu, Z., et al.: Swin transformer: Hierarchical vision transformer using shifted windows (2021), https://arxiv.org/abs/2103.14030
18. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
19. Marcinkevičs, R., et al.: Regensburg pediatric appendicitis dataset. Zenodo (2023), https://doi.org/10.5281/zenodo.7711412
20. Pedraza, L., et al.: An open access thyroid ultrasound image database. In: 10th International Symposium on Medical Information Processing and Analysis. vol. 9287, pp. 188–193. SPIE (2015)
21. Ronneberger, O., Fischer, P., Brox, et al.: U-net: Convolutional networks for biomedical image segmentation. In: Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv. (MICCAI). pp. 234–241. Springer (2015)
22. Sener, O., Koltun, V.: Multi-task learning as multi-objective optimization. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 31 (2018)
23. Shazeer, N.: Glu variants improve transformer. arXiv preprint arXiv:2002.05202 (2020)
24. Shazeer, N., et al.: Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In: International Conference on Learning Representations (ICLR) (2017)
25. Shen, D., Wu, G., Suk, H.I.: Deep learning in medical image analysis. Annual Review of Biomedical Engineering **19**, 221–248 (2017)
26. Singla, R., et al.: The open kidney ultrasound data set. In: International Workshop on Advances in Simplifying Medical Ultrasound. pp. 155–164. Springer (2023)
27. Wu, C., et al.: Dynamic modeling of patients, modalities and tasks via multi-modal multi-task mixture of experts. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=NJxCpMt0sf
28. Ye, H., Xu, D.: Taskexpert: Dynamically assembling multi-task representations with memorial mixture-of-experts. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 21771–21780 (2023). https://doi.org/10.1109/ICCV51070.2023.01995
29. Ye, J., et al.: Taskexpert: Dynamically assembling domain-specific experts for multi-task radiology report generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2137–2147 (2023)

30. Yu, T., et al.: Gradient surgery for multi-task learning. Advances in neural information processing systems **33**, 5824–5836 (2020)
31. Zhang, Y., Xian, M., Cheng, et al.: Busis: A benchmark for breast ultrasound image segmentation. Healthcare **10**(4), 729 (2022)