Towards Unified Image–Video Modeling Across B-Mode and Contrast-Enhanced Ultrasound for Multi-Task Analysis

Qiang Huang[1†], Dengqiang Jia[1†], Xiangmei Chen[2†], Zehui Lin[1], Zhizhou Wang[1], Ligang Cui[3], Zhiyuan Wang[4(✉)], and Tao Tan[1(✉)]

[1] Faculty of Applied Sciences, Macao Polytechnic University, Macao, China
[2] Department of Ultrasound, Peking University Shenzhen Hospital, Shenzhen, China
[3] Department of Ultrasound, Peking University Third Hospital, Beijing, China
[4] Department of Ultrasound, The Affiliated Cancer Hospital of Xiangya School of Medicine, Central South University, Changsha, China
[†] These authors contributed equally to this work.
[(✉)] Corresponding authors.
wangzhiyuan@hnca.org.cn; taotan@mpu.edu.mo

Abstract. Ultrasound imaging is widely used for disease diagnosis in clinical practice due to its non-invasive and portable advantages. However, current deep learning-based models are predominantly designed for specific organs or tasks, leading to performance with limited cross domain generalization and insufficient exploitation of available ultrasound data such as ultrasound images and videos. To address this limitation, we propose UIV-US, a unified multi-organ and multi-task framework for ultrasound image and video understanding across B-mode and contrast-enhanced ultrasound (CEUS). UIV-US operates within a shared parameter space to jointly support segmentation and classification tasks across both images and videos, enabling consistent representation learning across modalities and temporal settings. We design a Structured Semantic Conditioning Prompt (SSCP) that comprises five semantic attributes grouped into structural and spatial priors for adaptive conditioning of shared features, forming the foundation for scalable learning from large-scale ultrasound datasets. To further bridge image and video reasoning, a Memory-Reinforced Temporal Attention (MRTA) is introduced to ensure temporal consistency for video segmentation, while a segmentation-aware Feature-Level Attention (FLA) strategy performs reliable temporal aggregation for video classification. Extensive experiments on a large-scale multi-organ benchmark comprising 27,804 segmentation images, 7,320 classification images, and over 1,400 ultrasound videos demonstrate the effectiveness of UIV-US. The proposed framework achieves 89.81% Dice Similarity Coefficient for image segmentation, 96.28% for video segmentation, and 89.26% for CEUS segmentation. For classification, it reaches 85.76% Area Under the ROC Curve on image, 80.32% on video and 72.16% on CEUS, demonstrating strong performance across tasks. The code is available at https://github.com/miccai26-1426/UIV-US.

Keywords: Universal Framework · Image & Video Analysis · Multi-Task · Multi-Organ.

1 Introduction

Ultrasound imaging is one of the most widely used diagnostic tools in clinical practice and produces large-scale heterogeneous data across organs, modalities, and temporal settings [27]. A routine examination may simultaneously involve structural delineation, perfusion assessment, and diagnostic classification within a single session.

Despite the rapid progress of deep learning, most existing ultrasound analysis methods remain task- and organ-specific. Dedicated segmentation [14] and classification [7] networks dominate conventional B-mode applications, while CEUS approaches exploit contrast dynamics via cross-modality fusion [21] and temporal modeling [3]. For ultrasound videos, spatiotemporal aggregation [13] and lightweight temporal modules [23] are introduced for classification and segmentation. Although these methods achieve strong performance within restricted settings, image and video modeling are typically developed as separate pipelines, limiting parameter sharing and scalability.

To improve generalization, recent studies have explored multi-task and generalizable representation learning for ultrasound analysis. Large-scale pretraining frameworks such as USFM [10] and TinyUSFM [18] aim to learn transferable representations, and multi-organ segmentation strategies with shared representations and anatomical priors [4] have been investigated. In parallel, prompt-based paradigms establish flexible task conditioning mechanisms. Specifically, SAM [11] and SAM2 [25] introduce prompt-conditioned segmentation for images and videos, and medical adaptations including MedSAM [19], SAM-Med2D [5], and SAMUS [15] extend this strategy to clinical domains. More recent promptable frameworks such as UniUSNet [16] and PerceptGuide [17] support joint segmentation and classification for ultrasound images. However, most existing approaches remain predominantly image-oriented or segmentation-focused, and explicit unification of image and video reasoning across multiple ultrasound modalities within a single multi-task framework remains limited.

Motivated by these observations, we propose UIV-US, a unified multi-task image–video modeling framework for B-mode and contrast-enhanced ultrasound (CEUS). UIV-US operates within a shared parameter space to jointly support image and video segmentation and classification. To adapt shared representations to diverse clinical scenarios without increasing model complexity, we employ structured semantic conditioning and lightweight temporal reasoning within a single architecture.

Contributions: (1) We propose UIV-US, a shared-parameter ultrasound framework that unifies image and video segmentation and classification across B-mode and CEUS under a single architecture. (2) We design a Structured Semantic Conditioning Prompt (SSCP) mechanism that organizes five semantic attributes into structural and spatial priors to adaptively condition shared intermediate fea-

tures, providing a scalable foundation for learning from large-scale ultrasound datasets. (3) We introduce a temporal modeling scheme within the shared architecture: a Memory-Reinforced Temporal Attention (MRTA) module that enforces temporally consistent segmentation through dynamic memory encoding and cross-attention, and a Feature-Level Attention (FLA) strategy that enables segmentation-aware frame aggregation for video classification.

2 Method

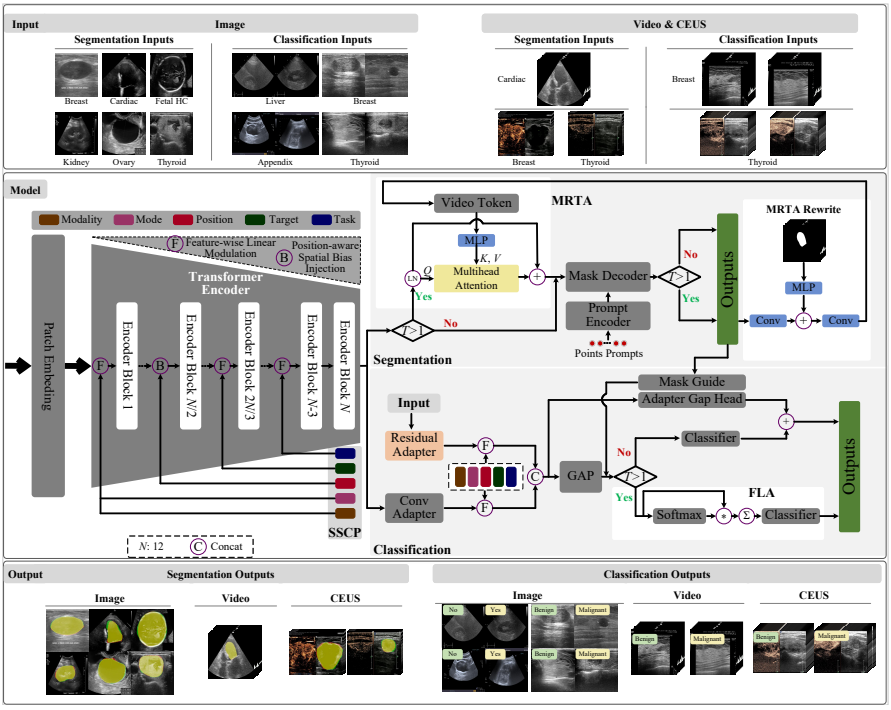

Fig. 1. The overview of UIV-US. The framework shares a Transformer encoder across segmentation and classification tasks, with SSCP for adaptive modulation. MRTA ensures temporal consistency for video segmentation, while FLA performs feature-level aggregation for video classification.

2.1 Overall Architecture

UIV-US is a universal multi-organ and multi-task model for unified ultrasound image and video understanding across B-mode and CEUS, as illustrated in Fig. 1. A single shared encoder is jointly optimized for segmentation and classification,

while SSCP inject structural and spatial priors to adaptively modulate the shared representation space. For video inputs, temporal consistency is modeled through MRTA for segmentation and FLA for classification, enabling unified spatiotemporal reasoning within the same parameter space.

2.2 Structured Semantic Conditioning Prompt (SSCP)

We introduce an SSCP mechanism that dynamically adapts the shared representation space without duplicating model parameters.

Given a sample, a structured semantic prompt is built from five attributes, which are organized into two types of priors. The structural priors (*Task*, *Target*, *Mode*, and *Modality*) are injected via learnable feature modulation, while *Position* is modeled as a spatial prior through spatial bias. These attributes are predefined categorical metadata obtained from task settings and dataset/examination labels and are available before model input. For each attribute $a \in \mathcal{A}$, a one-hot vector $\mathbf{p}_a$ is mapped through a learnable prompt bank as $\mathbf{e}_a = \mathbf{p}_a \mathbf{W}_a$. For the four structural attributes, $(\boldsymbol{\gamma}_a, \boldsymbol{\beta}_a) = \mathrm{MLP}_a(\mathbf{e}_a)$ modulates an intermediate feature as $\mathbf{h}' = \mathbf{h} \odot (1 + \tanh(\boldsymbol{\gamma}_a)) + \boldsymbol{\beta}_a$. For *Position*, we compute $\mathbf{c}_{\mathrm{pos}} = \mathrm{MLP}_{\mathrm{pos}}(\mathbf{e}_{\mathrm{pos}})$ and apply the spatial bias $\mathbf{h}'_{xy} = \mathbf{h}_{xy} + \mathbf{c}_{\mathrm{pos}} B_{xy}$, where $B_{xy} = (x + y)/2$ and $x, y \in [-1, 1]$.

Instead of concatenating prompt information to the input, SSCP modulates intermediate feature representations within the shared backbone. Different semantic components are injected at different depths of the network. Coarse-grained attributes such as modality and mode influence early layers, while task- and target-specific priors refine higher-level semantic features.

Importantly, SSCP introduces only lightweight conditioning layers, while backbone parameters remain shared across tasks and modalities. This design maintains a unified parameter space and enables flexible adaptation to diverse clinical scenarios without requiring task-specific model duplication. Since all attributes are predefined metadata, SSCP is deterministic and reproducible and requires no additional manual annotation at test time.

2.3 Spatiotemporal Reasoning and Unified Multi-Task Learning

For video inputs, segmentation adopts explicit memory-based temporal reasoning via Memory-Reinforced Temporal Attention (MRTA), whereas classification performs feature-level temporal aggregation using Feature-Level Attention (FLA) within the shared architecture.

Memory-Reinforced Temporal Attention (MRTA) for Video Segmentation. For video segmentation ($T > 1$), the standard encoder-decoder pipeline [11] is augmented with an MRTA to enforce temporal consistency across frames.

After predicting the segmentation mask $\hat{Y}_t$ for frame t, MRTA encodes a spatial memory representation

$$\mathbf{M}_t = \mathcal{E}_{\mathrm{mrta}}(\mathbf{F}_t, \hat{Y}_t), \tag{1}$$

where $\mathbf{F}_t$ denotes backbone features and $\mathcal{E}_{\text{mrta}}$ represents the MRTA encoder. The encoded memory is stored in a dynamic MRTA bank with K frames per video, containing conditioned reference frames and recent historical frames.

Before decoding frame $t+1$, the feature $\mathbf{F}_{t+1}$ is refined by MRTA:

$$\tilde{\mathbf{F}}_{t+1} = \mathcal{A}_{\text{mrta}}(\mathbf{F}_{t+1}, \{\mathbf{M}_{\leq t}\}), \tag{2}$$

where $\mathcal{A}_{\text{mrta}}$ denotes multi-head cross-attention over aggregated memory tokens derived from the encoded reference and recent frames. This refinement enhances inter-frame interaction and improves temporal stability without introducing recurrent networks or heavy video transformer architectures.

Feature-Level Attention (FLA) for Classification. Let $P = \{\mathbf{e}_a\}_{a \in \mathcal{A}}$ denote the SSCP embeddings. For image inputs ($T = 1$), classification operates on a segmentation-aware representation

$$\mathbf{z} = \mathcal{H}_{\text{feat}}(I, \tilde{\mathbf{F}}, \hat{Y}, P), \tag{3}$$

where global and mask-guided ROI representations are adaptively fused based on mask reliability. The final prediction is obtained via the shared classification head, i.e., $\hat{C} = \mathcal{H}_{\text{cls}}(\mathbf{z})$.

For video inputs ($T > 1$), frame-level representations $\{\mathbf{z}_t\}_{t=1}^{T}$ are first extracted using the same head:

$$\mathbf{z}_t = \mathcal{H}_{\text{feat}}(I_t, \tilde{\mathbf{F}}_t, \hat{Y}_t, P). \tag{4}$$

Feature-Level Attention (FLA) aggregates them into a video-level representation, where s_t denotes a learnable scalar importance score for frame t.

$$\mathbf{z}_{\text{video}} = \sum_{t=1}^{T} \alpha_t \mathbf{z}_t, \quad \alpha_t = \frac{\exp(s_t)}{\sum_{k=1}^{T} \exp(s_k)}, \tag{5}$$

which is then fed into the same classification head to produce video-level prediction, i.e., $\hat{C} = \mathcal{H}_{\text{cls}}(\mathbf{z}_{\text{video}})$.

3 Experiments

3.1 Experimental Setup

Experiments are conducted on a large-scale multi-organ ultrasound benchmark (UM^2-US) covering both B-mode and CEUS modalities. For image segmentation, we include BUS-BRA[8], TN3K [24], MMOTU [29], Fetal_HC [9], kidneyUS_capsule [26], and EchoNet-Dynamic [22], totaling 27,804 images. For image classification, BUS-BRA, TN3K, Fatty-Liver [2], and Appendix [20] are used, comprising 7,320 labeled images. Video segmentation is evaluated on CAMUS [12], while video classification uses a private breast ultrasound dataset

Table 1. Validation comparison of UIV-US with SOTA pretrained models on UM^2-US. We present the mean DSC over all segmentation datasets, and mean AUC/Accuracy over all classification datasets.

	Model	Task	Mode	Prompt	Seg(DSC↑)	Cls(AUC↑/Acc↑)
Image	SAM [11]	Seg	zero-shot	Box	71.23%	-
	SAM [11]		zero-shot	Point	37.13%	-
	SAM-Med2D [5]		zero-shot	Box	69.00%	-
	SAM-Med2D [5]		zero-shot	Point	31.24%	-
	SAM-Med2D [5]		fine-tune	Box	69.56%	-
	SAM-Med2D [5]		fine-tune	Point	72.95%	-
	MedSAM [19]		zero-shot	Box	78.21%	-
	MedSAM [19]		fine-tune	Box	67.25%	-
	SAMUS [15]		fine-tune	Point	85.91%	-
	UniUSNet [16]	Seg & Cls	-	One-hot	82.62%	78.64%/79.95%
	PerceptGuide [17]		-		87.62%	79.33%/80.69%
	Ours	All	-	Dual-Priors	**89.81%**	**85.76%/82.23%**
Video	SAMUS [15]	Seg	fine-tune	Point	92.35%	-
	MemSAM [6]	Seg	fine-tune	Point	95.23%	-
	Ours	All	-	Dual-Priors	**96.28%**	**80.32%/86.33%**
CEUS	**Ours**	All	-	Dual-Priors	**89.26%**	**72.16%/77.33%**

Breast* (300 videos). CEUS segmentation and classification are further evaluated on two private datasets, CEBreast* (300 videos) and CEThyroid* (300 videos).

All data are split at the patient level into 70%/10%/20% training, validation, and testing sets. The model is implemented in PyTorch and optimized using AdamW with an initial learning rate of 5×10^{-4} and weight decay of 0.001. Training is performed in two stages: segmentation is first optimized with the classification branch frozen, followed by classification training with the encoder and segmentation branch frozen. We employ Dice and cross-entropy losses for segmentation, and cross-entropy loss for classification.

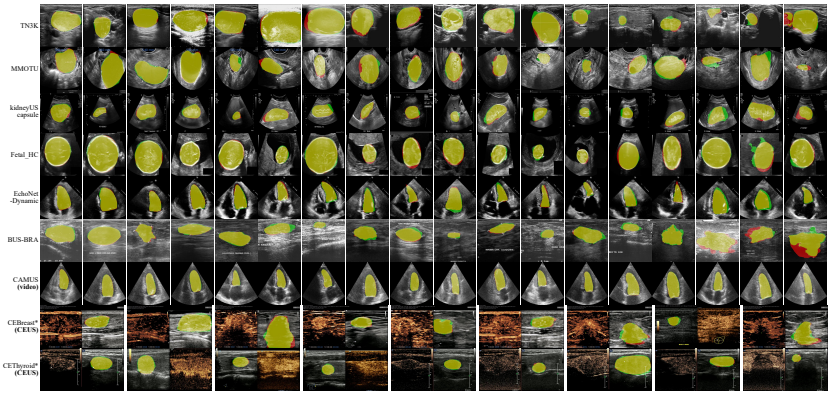

Fig. 2. Qualitative results of UIV-US across multiple datasets, with more accurate predictions on the left and slightly less precise results on the right. Yellow regions denote the overlap between prediction and ground truth, green and red regions denote false positives and false negatives, respectively.

Table 2. Ablation results of SSCP, MRTA, and FLA. DSC is reported for segmentation and AUC for classification. Inter-frame Dice (IF-D, ↑) and Area Variation (AV, ↓) measure temporal stability.

Task	Variant / Method	Image		Video		CEUS		Stability (Video)	
		DSC↑	AUC ↑	DSC↑	AUC↑	DSC↑	AUC ↑	IF-D ↑	AV ↓
All Task	W/o SSCP	87.34%	82.94%	94.72%	76.24%	86.51%	68.27%	–	–
	W/o FiLM	87.29%	84.81%	94.79%	77.26%	88.29%	71.39%	–	–
	W/o SpatialBias	89.15%	83.27%	95.50%	79.83%	89.02%	69.91%	–	–
	Ours	**89.81%**	**85.76%**	**96.28%**	**80.32%**	**89.26%**	**72.16%**	–	–
Video & CEUS Seg.	W/o MRTA	–	–	93.12%	–	83.81%	–	86.47%	0.2003
	MRTA(K=1)	–	–	94.49%	–	87.23%	–	95.84%	0.0529
	MRTA(K=3)	–	–	95.00%	–	86.62%	–	97.33%	0.0451
	MRTA(K=6)	–	–	95.43%	–	87.12%	–	97.49%	0.0410
	Ours(K=10)	–	–	**96.28%**	–	**89.26%**	–	**97.59%**	**0.0382**
Video & CEUS Cls.	Mean pooling	–	–	–	76.15%	–	69.83%	–	–
	Max pooling	–	–	–	74.51%	–	66.37%	–	–
	Transformer pooling	–	–	–	79.83%	–	71.24%	–	–
	FLA	–	–	–	**80.32%**	–	**72.16%**	–	–

3.2 Comparison with State-of-the-Art Methods

We conduct systematic comparisons across multiple ultrasound tasks to evaluate the overall performance of the proposed framework. As existing models are primarily designed for individual tasks, UIV-US is compared against task-specific state-of-the-art (SOTA) methods for each corresponding setting.

For image segmentation, we compare with representative SAM-based baselines, including SAM [11], MedSAM [19], SAM-Med2D [5], and SAMUS [15]. For video segmentation, we compare with video-adapted SAMUS and MedSAM, as well as the memory-based method MemSAM [6]. For image classification and multi-task learning, we further compare with unified frameworks such as UniUSNet [16] and PerceptGuide [17].

Table 1 presents quantitative comparisons with representative SOTA methods across image and video segmentation as well as classification tasks. For image segmentation, zero-shot SAM-based models achieve moderate performance. Fine-tuning improves performance but remains sensitive to prompt type. Domain-specific models such as SAMUS achieve stronger results, demonstrating the importance of task adaptation in ultrasound settings. For video segmentation, MemSAM achieve competitive DSC scores. In unified multi-task settings, existing approaches (UniUSNet and PerceptGuide) demonstrate promising joint learning capability for segmentation and classification.

In contrast, UIV-US consistently achieves superior performance across all tasks, reaching 89.81% DSC for image segmentation, 96.28% for video segmentation, and 89.26% for CEUS segmentation. Classification performance also improves to 85.76%/82.23% (AUC/Acc) for image tasks and 72.16%/77.33% for CEUS. Qualitative results are illustrated in Fig. 2. These results demonstrate the effectiveness of structured semantic prompting and unified optimization.

3.3 Ablation Study

To validate the effectiveness of SSCP, MRTA, and FLA, we remove each module individually under the same experimental setting. The results are presented in Table 2.

For SSCP, we evaluate three variants: w/o SSCP, w/o FiLM, and w/o SpatialBias. Removing SSCP degrades performance across all tasks, confirming the effectiveness of prompt-guided modulation. Both FiLM and SpatialBias contribute complementary improvements, and the full model achieves the best overall performance. For MRTA, we compare w/o MRTA and different memory lengths ($K = 1, 3, 6, 10$), and evaluate temporal stability using Inter-frame Dice (IFD) and Area Variation (AV). Removing MRTA significantly reduces video DSC (93.12%) and temporal stability (IF-D: 86.47%, AV: 0.2003). Increasing memory length consistently improves both accuracy and consistency, with the best performance achieved at $K = 10$. For FLA, we replace it with mean pooling, max pooling, and Transformer pooling. FLA achieves the best results, demonstrating the advantage of adaptive feature aggregation.

3.4 External Validation

To evaluate cross-domain generalization, we further conduct segmentation experiments on external ultrasound datasets, including BUSI [1], BUSIS [28], and DDTI [24], which are not used during training. UIV-US achieves DSC scores of 73.99%, 89.76%, and 67.83% on BUSI, BUSIS, and DDTI, respectively, maintaining competitive performance across domains. These results demonstrate the robustness and transferability of the proposed framework beyond the training datasets.

To examine the hierarchical evolution of feature representations, we analyze encoder and decoder features using kNN purity and centroid cosine distance. As shown in Fig. 3, encoder features exhibit stronger domain clustering, whereas decoder features show enhanced cross-domain alignment. This progressive change indicates a transition from domain-aware encoding to task-aligned decoding, suggesting reduced domain discrepancy in the decoder representation in the proposed framework.

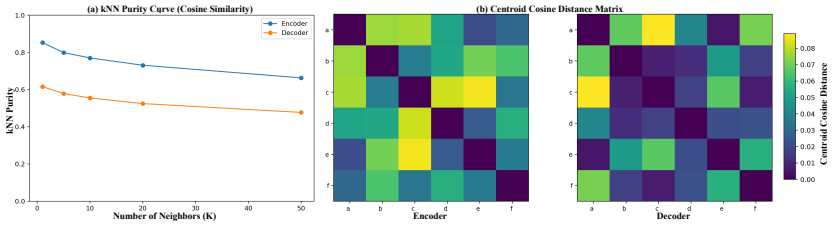

Fig. 3. Encoder-to-decoder feature evolution across all segmentation datasets. (a) kNN purity under cosine similarity for encoder and decoder representations across different neighborhood sizes K. (b) Inter-dataset centroid cosine distance matrices.

4 Conclusion

In this work, we present UIV-US, a unified multi-organ and multi-task framework for B-mode and CEUS ultrasound image and video understanding. Built upon a shared parameter space with structured and spatial semantic conditioning, the proposed model jointly supports segmentation and classification tasks across image and video within a single architecture. Through memory-reinforced temporal attention for video segmentation and feature-level attention for classification, UIV-US effectively exploits both spatial and temporal information without introducing task-specific model duplication. Extensive experiments on large-scale multi-organ benchmarks demonstrate consistent improvements over task-specific baselines, as well as generalization to three unseen external datasets. These results suggest that unified semantic-conditioned modeling provides an effective and scalable paradigm for comprehensive ultrasound analysis.

Acknowledgments. This work was supported by Shenzhen Medical Research Fund (D2501013), Macao Polytechnic University Grant (RP/FCA-13/2026), the Science and Technology Development Fund, Macau SAR (File no. 0004/2025/ASJ) under the FDCT-FAPESP Joint Funding Scheme, and the Science and Technology Development Fund of Macao under Grant 0041/2023/RIB2.

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., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data in brief **28**, 104863 (2020)
2. Byra, M., Styczynski, G., Szmigielski, C., Kalinowski, P., Michałowski, Ł., Paluszkiewicz, R., Ziarkiewicz-Wróblewska, B., Zieniewicz, K., Sobieraj, P., Nowicki, A.: Transfer learning with deep convolutional neural network for liver steatosis assessment in ultrasound images. International Journal of Computer Assisted Radiology and Surgery **13**, 1895–1903 (2018)
3. Chen, C., Wang, Y., Niu, J., Liu, X., Li, Q., Gong, X.: Domain knowledge powered deep learning for breast cancer diagnosis based on contrast-enhanced ultrasound videos. IEEE Transactions on Medical Imaging **40**(9), 2439–2451 (2021)
4. Chen, H., Cai, Y., Wang, C., Chen, L., Zhang, B., Han, H., Guo, Y., Ding, H., Zhang, Q.: Multi-organ foundation model for universal ultrasound image segmentation with task prompt and anatomical prior. IEEE Transactions on Medical Imaging **44**(2), 1005–1018 (2025)
5. Cheng, J., Ye, J., Deng, Z., Chen, J., Li, T., Wang, H., Su, Y., Huang, Z., Chen, J., Jiang, L., et al.: Sam-med2d. arXiv preprint arXiv:2308.16184 (2023)
6. Deng, X., Wu, H., Zeng, R., Qin, J.: Memsam: Taming segment anything model for echocardiography video segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9622–9631 (June 2024)

7. Ding, W., Wang, J., Zhou, W., Zhou, S., Chang, C., Shi, J.: Joint localization and classification of breast cancer in b-mode ultrasound imaging via collaborative learning with elastography. IEEE Journal of Biomedical and Health Informatics **26**(9), 4474–4485 (2022)
8. Gómez-Flores, W., Gregorio-Calas, M.J., Coelho de Albuquerque Pereira, W., J., M., Coelho de Albuquerque Pereira, W.: Bus-bra: A breast ultrasound dataset for assessing computer-aided diagnosis systems. Medical Physics (2023)
9. van den Heuvel, T.L., de Bruijn, D., de Korte, C.L., van Ginneken, B.: Automated measurement of fetal head circumference using 2d ultrasound images. PLoS One **13**(8), e0200412 (2018)
10. Jiao, J., Zhou, J., Li, X., Xia, M., Huang, Y., Huang, L., Wang, N., Zhang, X., Zhou, S., Wang, Y., Guo, Y.: Usfm: A universal ultrasound foundation model generalized to tasks and organs towards label efficient image analysis. Medical Image Analysis **96**, 103202 (2024)
11. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026. IEEE (2023)
12. 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.: Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. IEEE Transactions on Medical Imaging **38**(9), 2198–2210 (2019)
13. Li, M., Gong, W., Yan, P., Li, X., Jiang, Y., Luo, H., Zhou, H., Yin, S.: Joint lesion detection and classification of breast ultrasound video via a clinical knowledge-aware framework. IEEE Transactions on Circuits and Systems for Video Technology **35**(1), 45–61 (2025)
14. Li, R., He, Y., Ge, R., Wang, C., Zhang, D., Chen, Y., Li, S.: Gaze-assisted human-centric domain adaptation for cardiac ultrasound image segmentation. In: ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5 (2025)
15. Lin, X., Xiang, Y., Zhang, L., Yang, X., Yan, Z., Yu, L.: Samus: Adapting segment anything model for clinically-friendly and generalizable ultrasound image segmentation. arXiv preprint arXiv:2309.06824 (2023)
16. Lin, Z., Zhang, Z., Hu, X., et al.: Uniusnet: A promptable framework for universal ultrasound disease prediction and tissue segmentation. In: 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 3501–3504. IEEE (2024)
17. Lin, Z., Li, S., Wang, S., Gao, Z., Sun, Y., Lam, C.T., Hu, X., Yang, X., Ni, D., Tan, T.: An orchestration learning framework for ultrasound imaging: Prompt-guided hyper-perception and attention-matching downstream synchronization. Medical Image Analysis **104**, 103639 (2025)
18. Ma, C., Jiao, J., Liang, S., Fu, J., Wang, Q., Li, Z., Wang, Y., Guo, Y.: Tinyusfm: Towards compact and efficient ultrasound foundation models (2025), `https://api.semanticscholar.org/CorpusID:282272257`
19. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)
20. Marcinkevičs, R., Reis Wolfertstetter, P., Klimiene, U., Özkan Elsen, E., Chin-Cheong, K., Paschke, A., Vogt, J.E.: Regensburg pediatric appendicitis dataset (2023)

21. Meng, Z., Zhu, Y., Fan, X., Tian, J., Nie, F., Wang, K.: Ceusegnet: A cross-modality lesion segmentation network for contrast-enhanced ultrasound. In: 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI). pp. 1–5 (2022)
22. Ouyang, D., He, B., Ghorbani, A., Lungren, M.P., Ashley, E.A., Liang, D.H., Zou, J.Y.: Echonet-dynamic: A large new cardiac motion video data resource for medical machine learning. In: NeurIPS ML4H Workshop (2019)
23. Pang, X., Yao, F., Zhang, Y., Sun, Y., Patricio Lopes Lao, E., Lin, C., Cheong-Iao Pang, P., Wang, W., Li, W., Gao, Z., Tan, T.: Blenet: A bio-inspired lightweight and efficient network for left ventricle segmentation in echocardiography. IEEE Transactions on Circuits and Systems for Video Technology **35**(9), 9218–9233 (2025)
24. Pedraza, L., Vargas, C., Narváez, F., Durán, O., Muñoz, E., Romero, E.: An open access thyroid ultrasound image database. In: 10th International Symposium on Medical Information Processing and Analysis. vol. 9287, pp. 188–193. SPIE (2015)
25. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K.V., Carion, N., Wu, C.Y., Girshick, R., Dollár, P., Feichtenhofer, C.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
26. Singla, R., Ringstrom, C., Hu, G., Lessoway, V., Reid, J., Nguan, C., Rohling, R.: The open kidney ultrasound data set. In: International Workshop on Advances in Simplifying Medical Ultrasound. pp. 155–164. Springer (2023)
27. Zhang, T., Tan, T., Samperna, R., et al.: Radiomics and artificial intelligence in breast imaging: a survey. Artificial Intelligence Review **56**(Suppl 1), 857–892 (2023)
28. Zhang, Y., Xian, M., Cheng, H.D., Shareef, B., Ding, J., Xu, F., Huang, K., Zhang, B., Ning, C., Wang, Y.: Busis: A benchmark for breast ultrasound image segmentation. Healthcare **10**(4), 729 (2022)
29. Zhao, Q., Lyu, S., Bai, W., Cai, L., Liu, B., Cheng, G., Wu, M., Sang, X., Yang, M., Chen, L.: Mmotu: A multi-modality ovarian tumor ultrasound image dataset for unsupervised cross-domain semantic segmentation. arXiv preprint arXiv:2207.06799 (2022)