SAMASK-CLTR : A spatial-aware mask guided learning model for benign and malignant tumor classification in ABUS

Peirong Xu[1], Luoqian Zhu[2], Jingkun Chen[3], Xin Qian[1], Yue Sun[1], and Lingyun Bao[2] and Tao Tan[1*(✉)]

[1] Faculty of Applied Sciences, Macao Polytechnic University, Macao, China
taotan@mpu.edu.mo
[2] Hangzhou First People's Hospital, Hangzhou, China
[3] Department of Engineering Science, University of Oxford, Oxford, UK

Abstract. Automated Breast Ultrasound (ABUS) provides three dimensional volumetric imaging that improves breast lesion detection without radiation exposure and reduces operator dependency. However, the resulting high data volume poses significant challenges for radiologists in localizing lesions accurately and distinguishing benign from malignant cases—challenges that can directly impact early diagnosis and treatment outcomes. To tackle these critical issues, we propose SAMASK-CLTR (Spatial-Aware Mask Prompting with Convolutional Transformer Architecture), a hybrid framework that combines the feature extraction power of CNNs with the global modeling capability of Transformers. In our approach, ResNet-50 extracts hierarchical, multi-scale features that are refined by a Transformer encoder-decoder to capture global context. Crucially, during decoding, a mask prompt enhanced with 3D positional encoding guides the network to focus on key tumor regions, directly addressing the challenges of precise localization and classification. Experiments on 7,073 ABUS images—including 6,973 clinical cases from Internal Datasets and 100 cases from the public ABUS Challenge Cup—demonstrate that SAMASK-CLTR achieves AUCs of 88.45% and 70.46% on internal and external datasets, respectively. These results highlight the potential of our framework to significantly enhance breast cancer diagnosis by improving the accuracy and reliability of lesion classification. Code available at: https://github.com/SAMASK-CLTR/Code

Keywords: Auto Breast Ultrasound System · Computer Aided Diagnosis · Mask Prompt · Spatial Aware

1 Introduction

1.1 Background

Breast health is a critical concern for every woman. Among newly diagnosed female cancer cases, breast cancer accounts for up to 31% [21], making it the second leading cause of cancer death in women. Therefore, early detection of breast

* Corresponding author.

abnormalities, especially malignant changes, is crucial for increasing treatment success and improving quality of life, underscoring the importance of regular screenings [16].

Currently, three main methods are used for breast screening: mammography, breast ultrasound, and Automated Breast Ultrasound (ABUS). ABUS not only retains the advantages of traditional ultrasound but also captures three-dimensional ultrasonic information from multiple views, generating more comprehensive images of breast tissue [17]. This multi-view capability has increased its clinical adoption among both patients and doctors. During an ABUS examination, the device's probe automatically scans the entire breast, producing detailed three-dimensional images, which makes the process more convenient. Figure 1 shows the three views of ABUS. However, its multidimensional and multi-view nature also demands advanced interpretation skills from radiologists—especially junior physicians—as the large data volume can lead to higher rates of missed diagnoses. Although the application of Computer-Aided Detection (CAD) technology has improved diagnostic accuracy by assisting doctors in interpreting these complex images [27], the inherent challenges of 3D data remain.

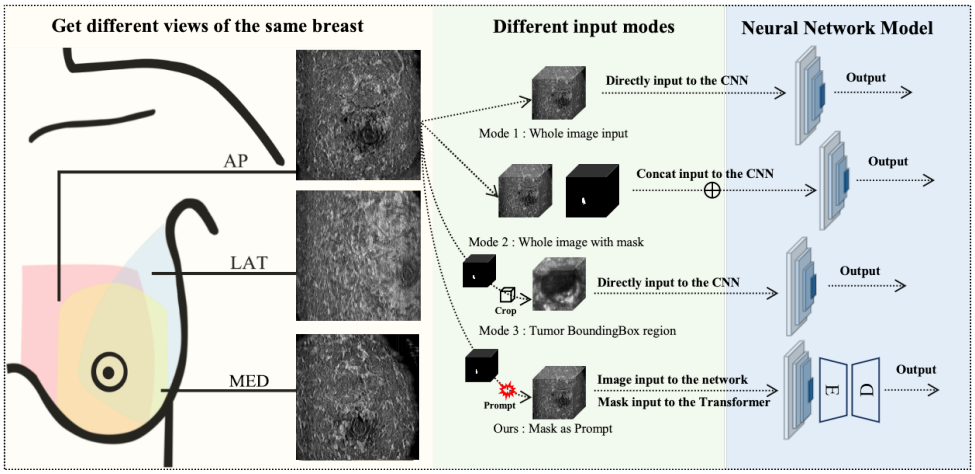

Fig. 1. This image shows different view of the same breast (AP, LAT, MED) and the framework of the four different input modes we compared.

In recent years, there has been remarkable progress in related tasks. For detection tasks, single-stage networks proposed in [29,26,14,19,24,18] are capable of fast and accurate lesion detection.For segmentation tasks, [2,3] suggest using a dynamic contrastive learning framework and a Teacher-Student Framework to improve segmentation performance. Meanwhile, models that incorporate prompt mechanisms, such as the Segment Anything Model (SAM) [9] and One-Prompt Segmentation [22], have demonstrated significant potential in the field of medical imaging. By leveraging prompt information [10,12,8,11,13,28,4,5], these models

significantly boost segmentation performance or exhibit strong adaptability for multi-task processing.

Although prompt-based models achieve significant performance improvements, their primary focus remains on the segmentation or localization of suspicious lesions, rather than directly determining the pathological nature of the lesions themselves. In benign and malignant classification tasks, many strategies aim to improve classification performance by operating on the region of interest (ROI). For instance, [32,1,20,25] reduce interference from surrounding redundant information, thereby enhancing classification accuracy. However, compared to ROI manipulation, directly utilizing prompts offers a more efficient and straightforward approach. Studies such as [23,7,6,15,30] adopt Vision Transformer (ViT) architectures for ultrasound image classification, but the global attention mechanism of ViT requires large amounts of data, which can easily lead to overfitting.

To overcome these challenges and directly address the pathological classification task, we propose **SAMASK-CLTR**, a spatial-aware mask-prompted classification framework for 3D breast ultrasound imaging. This framework employs ResNet-50 as the backbone for hierarchical feature extraction, combined with a Transformer to encode multi-scale semantic representations. A key innovation is the introduction of positionally encoded mask prompts during the decoding phase, which enhances the model's spatial awareness and tightly associates the feature space with tumor regions, significantly improving classification accuracy.The contributions of our research are threefold:

- We propose a hybrid CNN-Transformer model that integrates spatial-aware mask prompts for direct benign-malignant classification, achieving substantial improvements over conventional CNNs.
- We systematically evaluated multiple input modes and thoroughly analyzed their impact on classification performance.
- We conducted extensive experiments on large-scale clinical and public datasets, validating the cross-dataset generalization capability of SAMASK-CLTR.

2 Method

SAMASK-CLTR The SAMASK-CLTR model is a hybrid architecture that integrates convolutional neural networks with Transformer for three-dimensional data processing. Figure 2 shows the overview of our SAMASK-CLTR. The framework employs a 3D ResNet-50 network for multi-scale spatial feature extraction, where hierarchical feature maps are concatenated at the output stage and enhanced with three-dimensional positional encoding to capture spatial relationships. These encoded features are then fed into the Transformer encoder. In the decoder design, we replace the standard multi-head self-attention (MSA) mechanism with a Multi-Scale Deformable Attention module. This advanced mechanism adaptively samples features from multi-scale maps through learnable dynamic offsets, enabling the model to focus on task-relevant local regions while reducing computational cost and memory consumption, we integrate the query (Q) from the decoder with Spatial-aware Mask Prompts, combining tumor

mask information with dense spatial-scale encoding, and the Key (K) propagates feature information between the decoder layers. This fusion is then fed into the network to guide attention toward lesion areas more effectively. Finally, feature fusion from multiple decoder layers produces the prediction, with the entire network optimized end-to-end using cross-entropy loss for binary classification.

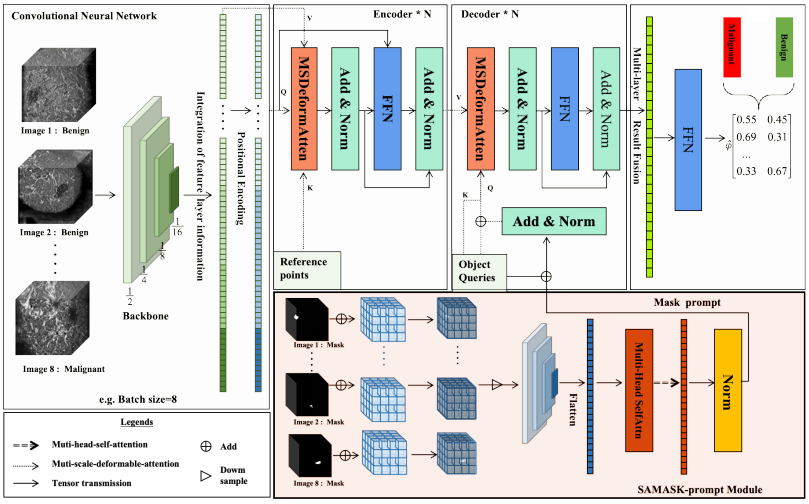

Fig. 2. SAMASK-CLTR: In this architecture, the backbone first extracts multi-scale feature information, and after applying positional encoding and flattening, feeds these features into the encoder module. Meanwhile, the Mask incorporates spatial positional encoding and downsampling before being fused with the decoder.

Deformable Attention and MSDeformable Attention Deformable Attention addresses the computational bottleneck of standard self-attention through dynamic sparse sampling [31]. Given a query element with feature z_q and reference point coordinates $\mathbf{p}_q \in \mathbb{R}^3$, the attention operation is formulated as:

$$\text{DeformAttn}(z_q, \mathbf{p}_q, x) = \sum_{m=1}^{M} W_m \left[\sum_{k=1}^{K} A_{mqk} \cdot W'_m x \left(\mathbf{p}_q + \Delta \mathbf{p}_{mqk} \right) \right], \quad (1)$$

where M denotes the number of attention heads (typically $M = 8$), K is the number of sampled points per head (empirically set to 4), $\Delta \mathbf{p}_{mqk} \in \mathbb{R}^3$ represents learnable offsets predicted from z_q, and A_{mqk} are attention weights normalized via softmax over K points. The trilinear interpolation operation $x(\cdot)$ enables differentiable sampling from feature volume x, facilitating gradient flow.

MSDeformable Attention extends this mechanism to multi-scale feature volumes $\{x_l\}_{l=1}^{L}$ with different spatial resolutions:

$$\text{MSDeformAttn}(z_q, \{\mathbf{p}_q^l\}) = \sum_{l=1}^{L}\sum_{k=1}^{K} A_{qkl} \cdot W_l x_l \left(\phi_l(\mathbf{p}_q) + \Delta\mathbf{p}_{qkl}\right). \tag{2}$$

Here $\phi_l(\cdot)$ performs coordinate scaling to match the l-th feature level's resolution (e.g., for $L = 4$ levels with downsampling rates of $\{1/2, 1/4, 1/8, 1/16\}$ along each axis relative to the input volume).

3D Position Embedding To extend positional information to volumetric data , we extend positional encoding from 2D to 3D to capture spatial relationships in three-dimensional space. We present the formula for 3D positional embedding here:

$$PE^{(z)}_{(pos,2i)} = \sin\left(\frac{z}{T^{\frac{2i}{d}}}\right), \quad PE^{(z)}_{(pos,2i+1)} = \cos\left(\frac{z}{T^{\frac{2i}{d}}}\right), \tag{3}$$

$$PE^{(y)}_{(pos,2i)} = \sin\left(\frac{y}{T^{\frac{2i}{d}}}\right), \quad PE^{(y)}_{(pos,2i+1)} = \cos\left(\frac{y}{T^{\frac{2i}{d}}}\right), \tag{4}$$

$$PE^{(x)}_{(pos,2i)} = \sin\left(\frac{x}{T^{\frac{2i}{d}}}\right), \quad PE^{(x)}_{(pos,2i+1)} = \cos\left(\frac{x}{T^{\frac{2i}{d}}}\right). \tag{5}$$

Spatial-aware Mask Prompts We propose a spatial-Aware mask prompt scheme. First, the tumor mask $M \in \mathbb{R}^{H\times W\times D}$, extracted from the input image, is enhanced via 3D sinusoidal-cosine positional encoding:

$$M_{\text{enc}}(x, y, z) = M(x, y, z) \oplus \left[PE_x(x) \parallel PE_y(y) \parallel PE_z(z)\right],$$

where PE_x, PE_y, PE_z are positional encoding functions along three axes, $\oplus$ denotes element-wise addition, and $\parallel$ represents channel concatenation. This ensures each voxel in the mask jointly encodes semantic information (tumor presence) and geometric 3D spatial coordinates. The encoded mask M_{enc} is then downsampled via multi-stage 3D convolutions to generate dimension-compatible features F, which are linearly combined with the decoder's initial query objects Q:

$$Q' = Q + F.$$

and fed into the decoder for iterative refinement. By explicitly modeling spatial-semantic correlations, this design significantly improves the model's localization accuracy and feature discriminability for tumor regions.

3 Experiment

3.1 Datasets

This study utilized an internal dataset along with the publicly available ABUS Challenge Cup external dataset. To ensure patient privacy, all data were anonymized,

with patient IDs replaced by unique identifiers. Tumor region masks were annotated by junior doctors invited by the research team, while bounding box images of tumor regions were generated by cropping based on the annotation files. The dataset was divided into an internal training set, an internal validation set, and an external validation set from the ABUS Challenge Cup dataset. This hierarchical partitioning strategy not only ensures rigorous model validation but also evaluates the model's generalization ability using independent external data, thereby supporting the universality and clinical applicability of the research findings. Table 1 shows the distribution of our datasets.

Table 1. Distribution of benign and malignant cases in internal and external datasets by patients, ABUS/ABVS images, mask annotations, and lesions (external data used for validation only)

Label	**Benign**	**Malignant**	**Total**
Patient	2259(+42 Externals)	941(+58 Externals)	3200(+100 Externals)
ABUS/ABVS	5501(+42 Externals)	1572(+58 Externals)	7073(+100 Externals)
MASK	3888(+42 Externals)	773(+58 Externals)	4661(+100 Externals)
Lesion	4947(+42 Externals)	779(+58 Externals)	5726(+100 Externals)

3.2 Experimental details

Our model using ResNet-50 as the backbone network for multi-scale feature extraction. Spatial features at 1/4, 1/8, and 1/16 resolutions were fused through cross-scale concatenation with 1×1 convolutions, then flattened and processed by a Transformer module containing 4 encoder-decoder layers (hidden dim 256, 6 attention heads). A set of 512 learnable object queries established feature correlations with the subsequent mask generation module, which incorporated a 3D spatial-aware positional encoding mechanism. The input volumes were preprocessed with random flipping (probability=0.3) and random contrast augmentation. The model was trained for 100 epochs with an initial learning rate of 1×10^{-4} , employing dynamic downsampling for query alignment and zero-initialized pseudo-masks to address annotation incompleteness.

In addition, we designed three input modes: Mode 1 directly inputs the whole image into the network; Mode 2 concatenates the image with its corresponding mask before inputting to the network; Mode 3 crops the lesion region from the image using the bounding box derived from the mask and inputs the cropped image into the network. Figure 1 shows illustrative diagrams of different input modes. All three modes, as well as our proposed model, adopted the same data augmentation techniques to ensure experimental consistency and comparability. Additionally, we addressed the issue of data imbalance by using random oversampling to increase the number of malignant samples.

3.3 Comparison Experiment

We systematically evaluated three input modes (Whole Image input, Image-Lesion Mask combined input, and Lesion Bounding Box Region input) and our proposed model on both internal and external datasets. Table 2 summarizes the performance metrics of different models under various input modes. The experimental results show that in the Whole Image input mode, ResNet101 achieved AUC values of 79.28% (internal) and 56.22% (external), demonstrating the best performance; in the Image-Lesion Mask combined input mode, DenseNet121 achieved AUC values of 76.45% (internal) and 62.71% (external), achieving the highest performance; and in the Lesion Bounding Box Region input mode, DenseNet121 achieved an AUC of 72.61% on the internal test set, while SwinUnetr achieved 55.42% on the external test set. Consequently, ResNet101, DenseNet121, and SwinUnetr were established as the state-of-the-art (SOTA) baseline models for their respective input modes. Meanwhile, we compared our proposed model with 3D DETR and SCPM-Net, and achieved AUC values of 88.45% on the internal validation set and 70.46% on the external independent validation set. A comprehensive analysis indicates that our method significantly outperforms existing models in terms of cross-dataset generalization and overall performance, with statistical significance ($p < 0.01$), further validating its robustness.

Table 2. Performance comparison of different modes on internal and external datasets

Method	Modes	Internal Dataset				External Dataset			
		AUC(%)	ACC(%)	SEN(%)	SPE(%)	AUC(%)	ACC(%)	SEN(%)	SPE(%)
Densenet121	W	76.75	83.19	54.45	<u>98.44</u>	49.74	49.44	35.14	69.41
	W+M	76.45	81.70	<u>58.48</u>	92.82	<u>62.71</u>	57.62	35.82	<u>87.97</u>
	B	72.61	75.39	51.76	90.13	49.23	49.55	33.03	67.39
Resnet18	W	76.61	83.54	53.59	97.89	53.64	55.48	43.09	72.73
	W+M	76.18	79.10	58.41	89.02	55.31	56.96	66.31	43.91
	B	72.33	75.72	49.86	86.12	50.45	55.39	44.89	62.32
Resnet50	W	76.60	81.40	55.90	93.62	49.09	56.84	71.88	36.12
	W+M	76.43	80.92	55.64	93.03	55.69	55.34	58.18	51.23
	B	69.31	74.54	53.72	84.53	45.33	52.76	59.32	42.76
Resnet101	W	<u>79.28</u>	<u>84.02</u>	54.45	98.20	56.22	56.36	<u>74.37</u>	28.31
	W+M	75.90	80.68	54.98	93.01	58.20	51.80	38.81	69.84
	B	67.93	77.76	51.13	87.92	52.75	52.33	34.29	66.14
SwimUnetr	W	78.48	82.99	53.63	97.07	48.87	49.44	30.43	75.87
	W+M	73.23	83.30	53.81	**98.92**	49.13	41.52	32.97	83.56
	B	67.22	72.43	49.88	90.34	55.42	<u>60.73</u>	22.42	71.31
3D DETR [26]	W	73.96	78.51	57.00	88.82	54.84	48.00	16.81	**91.14**
	B	68.29	72.88	47.77	82.76	52.17	45.56	24.18	88.72
SCPM-Net [14]	W	74.33	82.01	56.30	94.33	48.89	40.68	9.48	83.82
	B	66.73	80.13	53.29	87.38	45.75	41.54	13.43	82.14
Ours	W+SAMP	**88.45**	**84.13**	**79.63**	86.29	**70.46**	**70.90**	**76.68**	51.69

Mode abbreviations: W = Whole image, W+M = Whole image with mask, B = Lesion Bounding box region, W+SAMP = Whole image and spatial-aware mask prompt.

3.4 Ablation experiment

This study systematically evaluates three architectural configurations of the SAMASK-CLTR model through module ablation experiments: (1) baseline CLTR (without mask prompting module), (2) Mask Prompt CLTR (with mask prompting mechanism), and (3) Spatial-aware MASK prompt CLTR (joint optimization of mask prompting and positional encoding). Table 3 summarizes the results of the component ablation study on internal and external datasets. The experimental results demonstrate that the mask prompting module achieves an 2.21 percentage-point improvement in classification AUC compared to the baseline. The operates in two ways: During training, itguides feature learning by incorporating prior constraints on target regions, while during inference, it significantly enhances lesion localization accuracy. When jointly optimized with positional encoding, the model exhibits notable spatial awareness (11.43 percentage-point additional AUC gain). This dual-optimization strategy not only strengthens spatial feature representation but also effectively improves discriminative capability for capturing critical characteristics distinguishing benign and malignant lesions.

Table 3. Component Ablation Study on Internal and External Datasets

Components			Internal Datasets				External Datasets			
CLTR	Mask	Spatial	AUC%	ACC%	SEN%	SPE%	AUC%	ACC%	SEN%	SPE%
✓			77.02	77.20	60.95	85.00	62.55	56.60	59.89	50.08
✓	✓		79.23	80.87	62.32	82.76	63.94	62.41	62.74	50.33
✓	✓	✓	**88.45**	**84.13**	**79.63**	**86.29**	**70.46**	**70.90**	**76.68**	**51.69**

The '✓' mark indicates the modules that have been utilized.

4 Conclusion and Discussion

This study systematically analyzed the impact of different input modes on the classification performance of the model and proposed an improved spatially-aware mask prompt module to further enhance the model's performance. In the whole image input mode, the model can capture complete contextual information, but it is also easily defocused by background , which limits the improvement of classification performance. In the whole image with mask input mode, although the mask provides a rough localization of the tumor and theoretically helps the model focus on the target area, it fails to sufficiently suppress information from non-mask regions during training, which instead introduces additional interference and leads to a decline in performance. For the lesion region input mode, directly cropping the tumor region effectively reduces background interference, but at the same time, it loses critical contextual information, which limits the model's classification capability. This study innovatively introduces a spatially-aware mask prompting module. The proposed module combines 3D

positional encoding with the masked image, making the spatial contrast between the tumor region and normal tissue more pronounced. During training, this module guides the model to focus more accurately on the lesion region, suppressing noise information introduced by surrounding tissues, thereby achieving significant improvement in classification performance. Furthermore, this module demonstrates strong generalization ability in cross-dataset testing, further validating its effectiveness. Future work will focus on incorporating multimodal information prompts to further enhance the model's applicability in clinical settings.

Acknowledgments. This work is supported by Science and Technology Development Fund of Macao (0041/2023/RIB2), Macao Polytechnic University Grant (RP/FCA-05/2022).

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

References

1. Alotaibi, M., Aljouie, A., Alluhaidan, N., Qureshi, W., Almatar, H., Alduhayan, R., Alsomaie, B., Almazroa, A.: Breast cancer classification based on convolutional neural network and image fusion approaches using ultrasound images. Heliyon **9**(11) (2023)
2. Chen, J., Chen, C., Huang, W., Zhang, J., Debattista, K., Han, J.: Dynamic contrastive learning guided by class confidence and confusion degree for medical image segmentation. Pattern Recognition **145**, 109881 (2024)
3. Chen, J., Duan, H., Zhang, X., Gao, B., Tan, T., Grau, V., Han, J.: From gaze to insight: Bridging human visual attention and vision language model explanation for weakly-supervised medical image segmentation. arXiv preprint arXiv:2504.11368 (2025)
4. Chen, J., Huang, W., Zhang, J., Debattista, K., Han, J.: Addressing inconsistent labeling with cross image matching for scribble-based medical image segmentation. IEEE Transactions on Image Processing (2025)
5. Chen, J., Zhang, J., Debattista, K., Han, J.: Semi-supervised unpaired medical image segmentation through task-affinity consistency. IEEE Transactions on Medical Imaging **42**(3), 594–605 (2022)
6. Feng, H., Yang, B., Wang, J., Liu, M., Yin, L., Zheng, W., Yin, Z., Liu, C.: Identifying malignant breast ultrasound images using vit-patch. Applied Sciences **13**(6), 3489 (2023)
7. Gheflati, B., Rivaz, H.: Vision transformers for classification of breast ultrasound images. In: 2022 44th annual international conference of the EMBC. pp. 480–483. IEEE (2022)
8. Hu, M., Li, Y., Yang, X.: Breastsam: adapting the segmentation anything model for breast tumor segmentation in ultrasound imaging. In: Medical Imaging 2024: Ultrasonic Imaging and Tomography. vol. 12932, pp. 182–196. SPIE (2024)
9. 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 CVPR. pp. 4015–4026 (2023)

10. Li, H., Liu, H., Hu, D., Wang, J., Oguz, I.: Promise: Prompt-driven 3d medical image segmentation using pretrained image foundation models. In: ISBI. pp. 1–5. IEEE (2024)
11. Li, W., Liu, T., Feng, F., Yu, S., Wang, H., Sun, Y.: Btsspro: Prompt-guided multimodal co-learning for breast cancer tumor segmentation and survival prediction. IEEE Journal of Biomedical and Health Informatics (2024)
12. Li, X., Xu, M., Zhang, X., Niu, S., Zhu, J., Zhou, X.: Ceus-sam: Cross-modal prompt-based sam network for breast ceus image segmentation. In: 2024 BIBM. pp. 994–999. IEEE (2024)
13. Lin, Z., Zhang, Z., Hu, X., Gao, Z., Yang, X., Sun, Y., Ni, D., Tan, T.: Uniusnet: A promptable framework for universal ultrasound disease prediction and tissue segmentation. In: 2024 BIBM. pp. 3501–3504. IEEE (2024)
14. Luo, X., Song, T., Wang, G., Chen, J., Chen, Y., Li, K., Metaxas, D.N., Zhang, S.: Scpm-net: An anchor-free 3d lung nodule detection network using sphere representation and center points matching. Medical image analysis **75**, 102287 (2022)
15. Sun, J., Wu, B., Zhao, T., Gao, L., Xie, K., Lin, T., Sui, J., Li, X., Wu, X., Ni, X.: Classification for thyroid nodule using vit with contrastive learning in ultrasound images. Computers in biology and medicine **152**, 106444 (2023)
16. Sun, Y.S., Zhao, Z., Yang, Z.N., Xu, F., Lu, H.J., Zhu, Z.Y., Shi, W., Jiang, J., Yao, P.P., Zhu, H.P.: Risk factors and preventions of breast cancer. International journal of biological sciences **13**(11), 1387 (2017)
17. Tan, T., Platel, B., Mus, R., Karssemeijer, N.: Detection of breast cancer in automated 3d breast ultrasound. In: Medical Imaging 2012: Computer-Aided Diagnosis. vol. 8315, pp. 56–63. SPIE (2012)
18. Tang, J., Chen, X., Fan, L., Zhu, Z., Huang, C.: Ln-detr: An efficient transformer architecture for lung nodule detection with multi-scale feature fusion. Available at SSRN 5084519
19. Tao, X., Cao, Y., Jiang, Y., Wu, X., Yan, D., Xue, W., Zhuang, S., Yang, X., Huang, R., Zhang, J., et al.: Enhancing lesion detection in automated breast ultrasound using unsupervised multi-view contrastive learning with 3d detr. Medical Image Analysis p. 103466 (2025)
20. Wang, C., Guo, Y., Chen, H., Guo, Q., He, H., Chen, L., Zhang, Q.: Abus-net: Graph convolutional network with multi-scale features for breast cancer diagnosis using automated breast ultrasound. Expert Systems with Applications **273**, 126978 (2025)
21. Wen, X., Guo, X., Wang, S., Lu, Z., Zhang, Y.: Breast cancer diagnosis: A systematic review. Biocybernetics and Biomedical Engineering **44**(1), 119–148 (2024)
22. Wu, J., Xu, M.: One-prompt to segment all medical images. In: Proceedings of the CVPR. pp. 11302–11312 (2024)
23. Xu, M., Wang, W., Wang, K., Dong, S., Sun, P., Sun, J., Luo, G.: Vision transformers (vit) pretraining on 3d abus image and dual-capsvit: Enhancing vit decoding via dual-channel dynamic routing. In: 2023 BIBM. pp. 1596–1603. IEEE (2023)
24. Xu, Y., Shen, Y., Fernandez-Granda, C., Heacock, L., Geras, K.J.: Understanding differences in applying detr to natural and medical images. arXiv e-prints pp. arXiv–2405 (2024)
25. Yang, Z., Fan, T., Smedby, Ö., Moreno, R.: 3d breast ultrasound image classification using 2.5 d deep learning. In: 17th IWBI 2024. vol. 13174, pp. 443–449. SPIE (2024)
26. Zagoruyko, S.: End-to-end object detection with transformers. ECCV (2020)

27. van Zelst, J.C., Tan, T., Clauser, P., Domingo, A., Dorrius, M.D., Drieling, D., Golatta, M., Gras, F., de Jong, M., Pijnappel: Dedicated computer-aided detection software for automated 3d breast ultrasound; an efficient tool for the radiologist in supplemental screening of women with dense breasts. European radiology **28**, 2996–3006 (2018)
28. Zhao, Z., Zhang, Y., Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: One model to rule them all: Towards universal segmentation for medical images with text prompts. CoRR (2023)
29. Zhou, Y.T., Yang, T.Y., Han, X.H., Piao, J.C.: Thyroid-detr: Thyroid nodule detection model with transformer in ultrasound images. Biomedical Signal Processing and Control **98**, 106762 (2024)
30. Zhu, Q., Fei, L.: Cross-vit based benign and malignant classification of pulmonary nodules. PloS one **20**(2), e0318670 (2025)
31. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv e-prints pp. arXiv–2010 (2020)
32. Zhuang, Z., Ding, W., Zhuang, S., Raj, A.N.J., Wang, J., Zhou, W., Wei, C.: Tumor classification in automated breast ultrasound (abus) based on a modified extracting feature network. Computerized Medical Imaging and Graphics **90**, 101925 (2021)