MSQG-3DNet: Multi-Scale Token Selection with Query-Decoder and Gated Fusion for 3D Breast Tumor Classification

Zefeng Liu[1], Yaofei Duan[1], Zhikai Yang[2], Xiaojing Xu[3], Jie Gao[1], Jiaju Huang[1], Yue Sun[1], Kaiwen Yang[1], Lingyun Bao[3], and Tao Tan[(✉)1]

[1] Faculty of Applied Sciences, Macao Polytechnic University, Macao SAR
taotanjs@gmail.com
[2] Department of Biomedical Engineering and Health Systems, KTH Royal Institute of Technology, Stockholm, Sweden
[3] Department of Ultrasound, Affiliated Hangzhou First People's Hospital, School of Medicine, Westlake University, China

Abstract. Benign–malignant classification in automated breast ultrasound (ABUS) is challenging due to heterogeneous voxel spacing, anisotropic artifacts, and weakly localized malignancy cues distributed across 3D volumes. We propose MSQG-3DNet, a standardized-lattice CNN–Transformer hybrid for lesion-level ROI-based 3D ABUS classification, coupling multi-scale volumetric representation learning with query-guided global evidence aggregation. To reduce inter-case intensity and input-size variability, lesion-centered ROI volumes are intensity-normalized and standardized to a fixed $48 \times 48 \times 48$ lattice. The network employs an SE-enhanced dilated 3D residual encoder to capture multi-scale morphology, followed by dual-scale tokenization with hybrid 3D positional encoding, including coordinate-based encoding, sinusoidal encoding, and scale embedding, to preserve spatial and scale awareness under lattice standardization. LayerScale-stabilized Transformer blocks model long-range dependencies, and a lightweight cross-attention Query-Decoder extracts task-relevant evidence. Finally, a gated fusion integrates complementary global descriptors for malignancy prediction. On an in-house ABUS dataset (8,627 training and 859 test volumes), MSQG-3DNet achieves a test AUC of 96.23%. External evaluation on the public TDSC-ABUS benchmark yields the highest threshold-independent AUC of 71.55% among compared methods, indicating improved benign–malignant discrimination under domain shift, while highlighting the need for operating-threshold calibration in clinical use. Code available at: https://github.com/Ferry-Rain-Way/MSQG-3DNet

Keywords: Automated breast ultrasound · 3D tumor classification · CNN–Transformer hybrid

1 Introduction

Breast cancer is a major cause of cancer-related mortality among women worldwide [1], and early detection with accurate diagnosis remains a clinical prior-

ity [1]. In computer-aided diagnosis (CAD), reliable benign–malignant classification can reduce unnecessary follow-up examinations and invasive biopsies while improving the consistency of clinical decision-making. Nevertheless, despite the promise of AI-assisted breast imaging, real-world translation is often hindered by heterogeneous data distributions, variations in acquisition/workflow, and uneven evidence quality across studies [2].

In this study, we focus on lesion-level ROI-based benign–malignant classification after lesion localization, rather than end-to-end lesion detection or segmentation. In practical CAD workflows, the lesion ROI can be delineated by radiologists or generated by an upstream detection/segmentation module; MSQG-3DNet therefore serves as a downstream lesion characterization model. This lesion-level ROI-based setting is clinically meaningful as a downstream characterization task and is complementary to existing breast imaging CAD paradigms, including ROI-free ultrasound diagnosis, federated breast cancer diagnosis, and weakly supervised high-resolution breast screening models [22–24].

Automated breast ultrasound (ABUS) provides volumetric imaging that is attractive for screening and lesion assessment, yet benign–malignant classification in 3D ABUS is particularly challenging because (i) voxel spacing varies across cases and introduces geometric inconsistency, (ii) anisotropic artifacts degrade slice-wise continuity, and (iii) malignancy cues are frequently weak and dispersed across the 3D volume rather than being sharply localized. These characteristics demand models that can preserve fine-grained morphological details while also aggregating long-range evidence across slices.

3D convolutional neural networks (3D CNNs) are practical baselines for volumetric classification due to strong local inductive bias and effective modeling of voxel-level texture and structure. However, purely convolutional hierarchies rely on progressive downsampling to expand receptive fields, which can dilute subtle boundary irregularities and internal heterogeneity cues, and they remain limited in explicitly modeling long-range dependencies. Transformer-style global reasoning is a natural complement [3, 5], but direct application to 3D volumes is impeded by (i) limited annotated data and distributional heterogeneity, and (ii) prohibitive computation and memory when long token sequences are constructed from volumetric inputs [5, 7]. Moreover, naive tokenization of multi-scale volumetric features often introduces substantial background redundancy, further increasing cost and reducing discriminative focus.

To address these issues, we propose **MSQG-3DNet**, a progressively constructed CNN–Transformer hybrid for 3D breast tumor classification (Fig. 1). MSQG-3DNet first extracts stable local representations using an SE-MSAF 3D residual backbone. Multi-stage features are then tokenized with hybrid 3D positional/scale encoding, pruned to suppress background redundancy, encoded by Transformer blocks, aggregated by a Query-Decoder, and fused with complementary global descriptors through Gated Fusion. **Our contributions are three-fold:**

1. We propose a progressive CNN–Transformer hybrid architecture that couples local multi-scale enhancement with cross-scale token modeling.

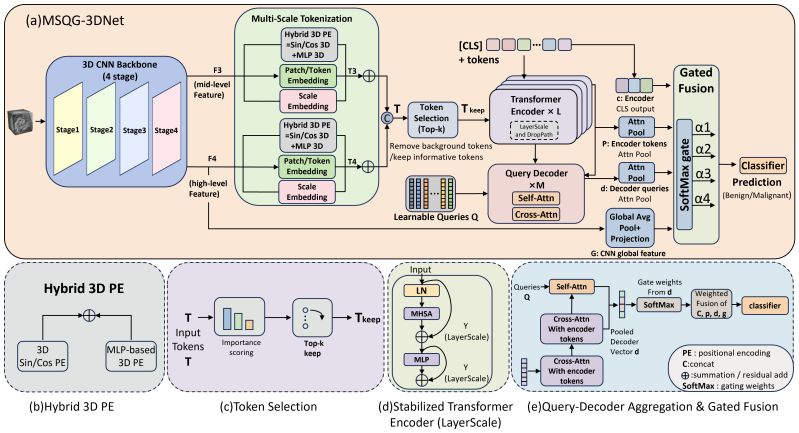

Fig. 1. Overview of MSQG-3DNet.

2. We design an efficient **Token Selection** → **Query-Decoder** aggregation pathway that reduces redundancy and strengthens task-focused global reasoning in 3D volumes.
3. We further introduce a **Gated Fusion** mechanism that integrates multi-branch global evidence for robust malignancy prediction, together with a reproducible 3D ABUS training pipeline.

2 Method

MSQG-3DNet follows a progressive lesion-level ROI-based design aligned with the ablation setting (Fig. 1). Given a lesion-centered ROI volume $X \in \mathbb{R}^{D\times H\times W}$, the model extracts multi-stage 3D features $\{F_i\}_{i\in\mathcal{I}}$, tokenizes and prunes them, encodes the selected tokens with Transformer blocks, aggregates evidence with a Query-Decoder, and fuses complementary descriptors to predict $\hat{y}$.

Multi-scale tokenization and positional encoding. We extract deep feature maps from multiple backbone stages and convert each feature map into a sequence of feature tokens:

$$T_i = \mathrm{Tok}(F_i) \in \mathbb{R}^{L_i\times E}, \qquad i \in \mathcal{I}, \tag{1}$$

where L_i denotes the number of tokens at stage i, and E is the embedding dimension. The multi-scale token sequences are concatenated as

$$T = \mathrm{Concat}(\{T_i\}_{i\in\mathcal{I}}). \tag{2}$$

To preserve spatial awareness after lattice standardization, we augment each token with a **hybrid 3D positional encoding** that combines coordinate-based encoding, sinusoidal encoding, and scale embedding:

$$\tilde{T} = T + P_{\mathrm{coord}} + P_{\mathrm{sin}} + P_{\mathrm{scale}}. \tag{3}$$

Since tokens from different scales are concatenated, the scale embedding prevents a scale-agnostic merger and enables spatial- and scale-aware Transformer interactions. Before Transformer encoding, the position-encoded token sequence $\tilde{T}$ is pruned by the Informative Token Pruning module described in Sec. 2.1, yielding the selected token sequence $\tilde{T}_s$. A learnable class token t_{cls} is then prepended to the selected sequence, which is fed into L Transformer blocks:

$$Z^{(0)} = [t_{\text{cls}}; \tilde{T}_s], \qquad Z^{(\ell)} = \text{TFM}(Z^{(\ell-1)}), \quad \ell = 1, \ldots, L. \tag{4}$$

The full sequence Z is used for query-guided aggregation, while Z_{tok} is pooled for fusion.

2.1 Informative Token Pruning

Multi-scale tokenization inevitably introduces substantial background redundancy, increasing computational cost and potentially diluting lesion-relevant cues. To address this issue, we introduce an **Informative Token Pruning** module to retain informative tokens before Transformer encoding.

Given the position-encoded token sequence $\tilde{T} = [\tilde{t}_1, \ldots, \tilde{t}_{L_T}]$, we estimate the importance of each token using a lightweight gating function:

$$a_j = \sigma(w_p^\top \tilde{t}_j + b_p), \qquad j = 1, \ldots, L_T, \tag{5}$$

where $\sigma(\cdot)$ denotes the sigmoid function. We then retain the top-K tokens according to the scores $\{a_j\}$:

$$\tilde{T}_s = \text{Select}(\tilde{T}, \{a_j\}_{j=1}^{L_T}, K), \tag{6}$$

where K is determined as a fixed proportion of the token sequence length. This operation reduces the computational burden of the subsequent Transformer encoder and encourages the model to focus on lesion-relevant evidence before global dependency modeling.

2.2 Query-Guided Evidence Aggregation

Simple token pooling provides only a coarse global summary and may be insufficient when malignancy-related cues are sparse, subtle, or spatially distributed. We therefore introduce a **Query-Guided Evidence Aggregation** module to perform task-oriented evidence extraction via cross-attention.

Let $Q^{(0)} \in \mathbb{R}^{M \times E}$ denote a set of M learnable queries. The decoder iteratively updates these queries by attending to the encoded selected sequence Z:

$$Q^{(\ell)} = \text{Dec}(Q^{(\ell-1)}, Z), \qquad \ell = 1, \ldots, L_D, \tag{7}$$

where each decoder block consists of self-attention among queries and cross-attention from queries to the encoded sequence. The resulting query representations $Q = Q^{(L_D)}$ are pooled to obtain a query-derived global descriptor:

$$z_{\text{qry}} = \text{Pool}(Q). \tag{8}$$

This mechanism explicitly allocates representational capacity to capture diverse evidence patterns, providing a more targeted and stable global representation than relying solely on the class token or naive token pooling.

2.3 Adaptive Multi-Descriptor Fusion

Different global descriptors capture complementary aspects of the input volume, but their reliability may vary across cases. We therefore employ an **Adaptive Multi-Descriptor Fusion** module to integrate multiple global representations in a sample-adaptive manner and reduce dependence on any single decision pathway.

Specifically, we construct four descriptors: (i) the Transformer class token z_{cls}, (ii) a convolutional global descriptor z_{conv} obtained by global pooling over deep backbone features, (iii) a token-pooled descriptor $z_{\text{tok}} = \text{Pool}(Z_{\text{tok}})$, and (iv) the query-derived descriptor z_{qry}. Fusion weights are predicted dynamically and the fused representation is formed as

$$\mathbf{w} = \text{softmax}(W_g z_{\text{qry}} + b_g) \in \mathbb{R}^4, \qquad z_f = \text{Concat}(\{w_k z_k\}_{k=1}^4), \tag{9}$$

where $\{z_k\} = \{z_{\text{cls}}, z_{\text{conv}}, z_{\text{tok}}, z_{\text{qry}}\}$. The final malignancy probability is then given by

$$\hat{y} = \text{sigmoid}(W z_f + b). \tag{10}$$

2.4 Loss Function

We train MSQG-3DNet with a hybrid objective to improve learning under class imbalance. The total loss combines cross-entropy loss and focal loss:

$$\mathcal{L} = (1 - \lambda)\mathcal{L}_{\text{CE}} + \lambda \mathcal{L}_{\text{Focal}}, \tag{11}$$

where

$$\mathcal{L}_{\text{Focal}} = -\alpha_y (1 - p_y)^\gamma \log(p_y). \tag{12}$$

Here, p_y denotes the predicted probability of the ground-truth class y, α_y is the class-specific weighting factor, and γ is the focusing parameter. This objective emphasizes hard examples while maintaining stable optimization for volumetric classification.

3 Experiment

Datasets. We evaluate MSQG-3DNet on both an in-house ABUS dataset and the public TDSC-ABUS dataset released in the MICCAI 2023 Challenge [8]. Dataset statistics are summarized in Table 1. All data were strictly de-identified in accordance with institutional review requirements.

Table 1. Statistics of the in-house and external TDSC-ABUS datasets.

Dataset	Split	Malignant	Benign	Total
In-house	Train	1336	7291	8627
In-house	Test	347	512	859
TDSC-ABUS	Train	87	64	151
TDSC-ABUS	Test	29	20	49

In-house ABUS dataset. The in-house dataset contains 9,486 ABUS volumes with lesion annotations. Tumor masks were first delineated by trained junior radiologists and subsequently verified by senior reviewers. Based on these annotations, lesion-centered 3D bounding-box crops were generated programmatically. The dataset was split into 8,627 cases for training and 859 cases for testing.

External TDSC-ABUS dataset. To assess out-of-distribution generalization, we further evaluate on the public TDSC-ABUS dataset [8], which contains 200 volumes (151 for training and 49 for testing). Public 3D ABUS datasets for lesion-level classification remain scarce; to our knowledge, TDSC-ABUS2023 is the only suitable public challenge benchmark and has been used by SAMASK-CLTR [9]. Its different acquisition/annotation setting and missing voxel-spacing metadata make it suitable for assessing relative robustness under domain shift.

Data Preprocessing. To reduce intensity and geometric variation across scans, each lesion-centered ROI volume was normalized and standardized before training. Specifically, truncated Z-score normalization was applied to suppress the influence of extreme intensity outliers. Because TDSC-ABUS2023 lacks voxel-spacing metadata, precise physical-scale normalization is unavailable for this external benchmark. Thus, all ROIs were interpolated to a fixed $48 \times 48 \times 48$ lattice for size standardization.

Evaluation Metrics. We report area under the ROC curve (AUC), accuracy (ACC), sensitivity (SEN), and specificity (SPE). SEN and SPE are defined as

$$\mathrm{SEN} = \frac{TP}{TP + FN}, \qquad \mathrm{SPE} = \frac{TN}{TN + FP}. \tag{13}$$

Unless otherwise specified, ACC, SEN, and SPE are computed using the default probability threshold of 0.5. Because clinical deployment may require different false-negative and false-positive trade-offs, we additionally analyze the operating point selected by Youden's index, $J = \mathrm{SEN} + \mathrm{SPE} - 1$, on the ROC curve.

Implementation Details. All models are trained under identical data splits and optimization settings for fair comparison. Training is performed for 100 epochs using the hybrid loss defined in Sec. 2.4. Class imbalance is addressed using class weighting and weighted sampling.

Computational efficiency. For a $48 \times 48 \times 48$ ROI input, MSQG-3DNet contains 165.91M trainable parameters and requires 7.55 GMACs, corresponding to 15.10 GFLOPs, per forward pass. The model is designed for accurate offline lesion characterization rather than real-time full-volume screening. The lesion-centered ROI setting and informative token pruning are intended to reduce redundant background computation before query-guided evidence aggregation.

Table 2. Progressive ablation results on the in-house test set. The variants progressively introduce the residual backbone, MSAF, and Transformer-based global reasoning modules. Best results are shown in bold.

Version	Residual Backbone	MSAF	TFM	AUC (%)	ACC (%)	SEN (%)	SPE (%)
V1	–	–	–	86.35	80.21	51.01	100.00
V2	✓	–	–	95.10	91.15	78.39	99.80
V3	✓	✓	–	95.37	91.04	**81.56**	96.68
MSQG-3DNet	✓	✓	✓	**96.23**	**91.27**	78.39	**100.00**

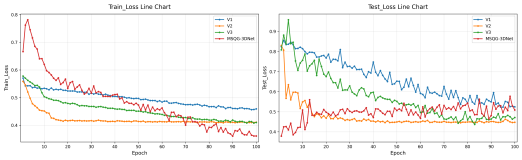

Fig. 2. Training and test loss curves of different model variants on the in-house dataset. The proposed MSQG-3DNet converges to a lower and more stable test loss than the baseline variants.

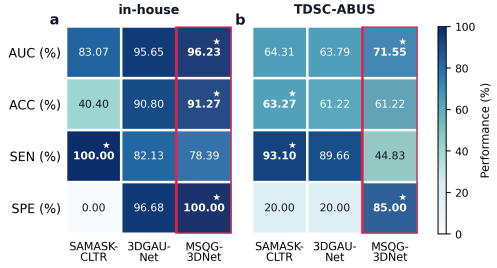

Fig. 3. Heatmap comparison of AUC, ACC, SEN, and SPE across datasets. MSQG-3DNet achieves the highest AUC on both datasets and shows markedly improved specificity on the external TDSC-ABUS cohort at the default threshold.

Ablation Study. We first analyze the contribution of each architectural component using the progressive design in Table 2. Starting from the plain baseline V1, introducing the residual 3D backbone (V2) yields a substantial gain in AUC from 86.35% to 95.10% and in ACC from 80.21% to 91.15%, indicating that stable hierarchical feature propagation is essential for volumetric ABUS classification. Adding MSAF (V3) further improves AUC to 95.37% and increases SEN from 78.39% to 81.56%, suggesting that multi-scale feature enhancement is beneficial for capturing lesion morphology and weak malignant patterns. Finally, the full MSQG-3DNet achieves the best AUC (96.23%) and ACC (91.27%), together with 100.00% specificity at the default threshold. These results indicate that the Transformer-based global reasoning pathway provides complementary discriminative information beyond purely convolutional modeling.

Table 3. Comparison with competing methods on the in-house test set. Best results are shown in bold.

Metric	SAMASK-CLTR	3DGAUnet	MSQG-3DNet
AUC (%)	83.07	95.65	**96.23**
ACC (%)	40.40	90.80	**91.27**
SEN (%)	**100.00**	82.13	78.39
SPE (%)	0.00	96.68	**100.00**

Table 4. Comparison with competing methods on the external TDSC-ABUS test set. Best results are shown in bold.

Metric	SAMASK-CLTR	3DGAUnet	MSQG-3DNet
AUC (%)	64.31	63.79	**71.55**
ACC (%)	**63.27**	61.22	61.22
SEN (%)	**93.10**	89.66	44.83
SPE (%)	20.00	20.00	**85.00**

Training dynamics. The training curves in Fig. 2 show that MSQG-3DNet converges to a lower and smoother test loss than earlier variants, suggesting improved optimization stability.

Comparison with Competing Methods. Given the scarcity of publicly reproducible 3D ABUS classification baselines, we compare MSQG-3DNet with SAMASK-CLTR [9] and 3DGAUnet [10], a representative 3D volumetric baseline adapted to the same classification setting, under the same ROI-based data split, preprocessing, loss function, and optimization settings. **In-house dataset.** As shown in Table 3, MSQG-3DNet achieves the best AUC (96.23%) and ACC (91.27%) among all methods. At the default threshold, it obtains 100.00% specificity, indicating strong false-positive suppression. Compared with 3DGAUnet, the proposed model further improves both AUC and specificity, demonstrating the benefit of global token-based reasoning for 3D ABUS classification. **External dataset.** On the external TDSC-ABUS dataset (Table 4), MSQG-3DNet achieves the highest AUC (71.55%), outperforming SAMASK-CLTR and 3DGAUnet by 7.24 and 7.76 percentage points, respectively. At the default threshold of 0.5, MSQG-3DNet favors false-positive suppression, achieving the highest SPE but lower SEN. With Youden's index, it obtains a more balanced SEN/SPE of 68.97%/60.00%, indicating that AUC reflects threshold-independent discrimination, whereas SEN and SPE depend on the clinical operating point.

Visualization Analysis. Fig. 3 provides a visual summary of the quantitative results. MSQG-3DNet achieves the strongest AUC/ACC profile on the in-house dataset and the highest AUC on the external dataset, with a clear specificity advantage at the default threshold.

4 Conclusion

We presented **MSQG-3DNet** for lesion-level ROI-based benign–malignant classification in 3D ABUS, targeting the practical challenges of geometric heterogeneity, anisotropic artifacts, and weak malignancy cues distributed across volumes. The proposed CNN–Transformer hybrid combines multi-scale volumetric feature learning with cross-scale token modeling, and introduces Token Selection, a Query-Decoder for task-driven global evidence aggregation, and Gated Fusion to integrate complementary global descriptors for stable prediction. MSQG-3DNet achieved a test AUC of **96.23%** on the in-house dataset and an AUC of **71.55%** on the external TDSC-ABUS dataset. Future work will focus on multi-center validation with complete voxel-spacing metadata, clinical operating-threshold calibration, domain-robust learning, and integration with upstream detection or segmentation modules toward an end-to-end CAD workflow.

Acknowledgments. This work was supported by Shenzhen Medical Research Fund (D2501013), the Science and Technology Development Fund of Macau SAR (File no. 0009/2025/itp1) 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. World Health Organization (WHO): Breast cancer. WHO Fact Sheet (14 August 2025). https://www.who.int/news-room/fact-sheets/detail/breast-cancer (last accessed: 25 February 2026).
2. Abdullah, K.A., Marziali, S., Nanaa, M., Escudero Sánchez, L., Payne, N.R., Gilbert, F.J.: Deep learning-based breast cancer diagnosis in breast MRI: systematic review and meta-analysis. European Radiology **35**, 4474–4489 (2025). https://doi.org/10.1007/s00330-025-11406-6
3. Brahmareddy, A., Selvan, M.P.: TransBreastNet a CNN transformer hybrid deep learning framework for breast cancer subtype classification and temporal lesion progression analysis. Scientific Reports **15**, 35106 (2025). https://doi.org/10.1038/s41598-025-19173-6
4. Jiang, C., Wang, Y., Yuan, Q., Qu, P., Li, H.: A 3D medical image segmentation network based on gated attention blocks and dual-scale cross-attention mechanism. Scientific Reports **15**, 6159 (2025). https://doi.org/10.1038/s41598-025-90339-y
5. Liu, S., Wang, L., Yue, W.: An efficient medical image classification network based on multi-branch CNN, token grouping Transformer and mixer MLP. Applied Soft Computing **153**, 111323 (2024). https://doi.org/10.1016/j.asoc.2024.111323
6. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable DETR: Deformable Transformers for End-to-End Object Detection. In: International Conference on Learning Representations, ICLR (2021). https://openreview.net/forum?id=gZ9hCDWe6ke

7. Marchetti, M., Traini, D., Ursino, D., Virgili, L.: Efficient token pruning in Vision Transformers using an attention-based Multilayer Network. Expert Systems with Applications **279**, 127449 (2025). https://doi.org/10.1016/j.eswa.2025.127449
8. Luo, G., Xu, M., Chen, H., Liang, X., Tao, X., Ni, D., et al.: Tumor Detection, Segmentation and Classification Challenge on Automated 3D Breast Ultrasound: The TDSC-ABUS Challenge. arXiv:2501.15588 (2025). https://doi.org/10.48550/arXiv.2501.15588
9. Xu, P., Zhu, L., Chen, J., Qian, X., Sun, Y., Bao, L., Tan, T.: SAMASK-CLTR: A spatial-aware mask guided learning model for benign and malignant tumor classification in ABUS. In: Gee, J.C., Hong, J., Sudre, C.H., Golland, P., Alexander, D.C., Iglesias, J.E., Venkataraman, A., Kim, J.H. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2025. LNCS, vol. 15960, pp. 567–577. Springer, Cham (2026). https://doi.org/10.1007/978-3-032-04927-8_54
10. Shi, Y., Tang, H., Baine, M.J., Hollingsworth, M.A., Du, H., Zheng, D., Zhang, C., Yu, H.: 3DGAUnet: 3D Generative Adversarial Networks with a 3D U-Net Based Generator to Achieve the Accurate and Effective Synthesis of Clinical Tumor Image Data for Pancreatic Cancer. Cancers **15**(23), 5496 (2023). https://doi.org/10.3390/cancers15235496
11. Wu, Y., Chaddad, A.: Impact of domain adaptation in deep learning for medical image classifications. arXiv:2602.09355 (2026). https://doi.org/10.48550/arXiv.2602.09355
12. Tan, T., Platel, B., Mus, R., Karssemeijer, N.: Detection of breast cancer in automated 3D breast ultrasound. In: Medical Imaging 2012: Computer-Aided Diagnosis. Proc. SPIE, vol. 8315, 831505 (2012). https://doi.org/10.1117/12.911068
13. 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), 1387–1397 (2017). https://doi.org/10.7150/ijbs.21635
14. van Zelst, J.C.M., Tan, T., Clauser, P., Domingo, A., Dorrius, M.D., Drieling, D., Golatta, M., Gras, F., de Jong, M., Pijnappel, R., Rutten, M.J.C.M., Karssemeijer, N., Mann, R.M.: 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**(7), 2996–3006 (2018). https://doi.org/10.1007/s00330-017-5280-3
15. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-End Object Detection with Transformers. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020. LNCS, vol. 12346, pp. 213–229. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58452-8_13
16. Tao, X., Cao, Y., Jiang, Y., Wu, X., Yan, D., Xue, W., Zhuang, S., Yang, X., Huang, R., Zhang, J., Ni, D.: Enhancing lesion detection in automated breast ultrasound using unsupervised multi-view contrastive learning with 3D DETR. Medical Image Analysis **101**, 103466 (2025). https://doi.org/10.1016/j.media.2025.103466
17. 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). https://doi.org/10.1016/j.eswa.2025.126978
18. Zhu, Q., Fei, L.: Cross-ViT based benign and malignant classification of pulmonary nodules. PLOS ONE **20**(2), e0318670 (2025). https://doi.org/10.1371/journal.pone.0318670

19. 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). https://doi.org/10.1016/j.bbe.2024.01.002
20. 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:2504.11368 (2025). https://doi.org/10.48550/arXiv.2504.11368
21. 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 IEEE International Conference on Bioinformatics and Biomedicine, BIBM, pp. 3501–3504. IEEE (2024). https://doi.org/10.1109/BIBM62325.2024.10822429
22. Mo, Y., Han, C., Liu, Y., Liu, M., Shi, Z., Lin, J., Zhao, B., Huang, C., Qiu, B., Cui, Y., Wu, L., Pan, X., Xu, Z., Huang, X., Li, Z., Liu, Z., Wang, Y., Liang, C.: HoVer-Trans: Anatomy-aware HoVer-Transformer for ROI-free breast cancer diagnosis in ultrasound images. IEEE Transactions on Medical Imaging **42**(6), 1696–1706 (2023). https://doi.org/10.1109/TMI.2023.3236011
23. Deng, T., Huang, C., Cai, M., Liu, Y., Liu, M., Lin, J., Shi, Z., Zhao, B., Huang, J., Liang, C., Han, G., Liu, Z., Wang, Y., Han, C.: FedBCD: Federated Ultrasound Video and Image Joint Learning for Breast Cancer Diagnosis. IEEE Transactions on Medical Imaging **44**(6), 2395–2407 (2025). https://doi.org/10.1109/TMI.2025.3532474
24. Shen, Y., Wu, N., Phang, J., Park, J., Liu, K., Tyagi, S., Heacock, L., Kim, S.G., Moy, L., Cho, K., Geras, K.J.: An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization. Medical Image Analysis **68**, 101908 (2021). https://doi.org/10.1016/j.media.2020.101908