PHYJOINTNET: Attention-Guided Joint Segmentation & Classification of Breast Ultrasound with Latent Feature Augmentation

Nimalesh Elangovan, Tushar Shinde, Hitika Tiwari
IIT Madras Zanzibar, Tanzania
zda24m008iitmz.ac.in, shindeiitmz.ac.in, hitika@iitmz.ac.in

Abstract

Ultrasound imaging is widely used for early breast cancer detection, but building automated systems that can both localize segmentation and assess classification remains challenging. These challenges stem from task interference, class imbalance in clinical datasets, and variability in image appearance due to differences in ultrasound hardware. To address this, we propose PhysJointNet, a physics-based augmentation framework for joint segmentation and classification in breast ultrasound. The framework enhances robustness and generalization across imaging devices while supporting effective multi-task learning. It includes a physics-informed augmentation module that uses a frozen descriptor to capture and align scanner-specific characteristics, reducing hardware-induced variations in texture and signal gain. Experimental results on benchmark datasets show improved performance, achieving 0.853 in segmentation and 0.918 in classification, demonstrating the effectiveness of the proposed approach for reliable clinical screening.

1. Introduction and Related Work

Breast cancer remains a major global health concern, where early detection is critical for improving treatment outcomes and patient survival rates. Ultrasound imaging is widely adopted for screening due to its non-invasive nature, absence of ionizing radiation, and suitability for dense breast tissues and diverse patient populations [3]. However, ultrasound interpretation is inherently challenging due to variations in scanner hardware, acquisition settings, complex anatomical structures, and subtle pathological cues, leading to diagnostic variability [15]. These challenges motivate the development of automated systems for lesion localization and malignancy assessment.

Deep learning methods have achieved strong performance in either segmentation or classification [9], but most approaches treat them independently, overlooking their interdependence in clinical decision-making [10]. Joint learning frameworks improve feature sharing [16, 20], yet often suffer from task interference, where competing objectives hinder performance [4]. Additionally, breast ultrasound datasets exhibit domain shifts due to scanner-specific characteristics such as speckle patterns, signal attenuation, and gain, which are not well captured by conventional geometric augmentations [18].

Classification methods using transfer learning and ensembles struggle with domain shifts and class imbalance [12], while segmentation approaches based on U-Net variants [19] excel in localization but lack diagnostic integration. GAN-based augmentation has been explored [22], but often fails to preserve fine anatomical structures. Neural Style Transfer using AdaIN [11] provides a stable alternative for modeling appearance variations, though its use in medical imaging remains limited. Furthermore, increasing model complexity limits deployment on resource-constrained point-of-care devices.

To address these challenges, we propose PHYSJOINTNET, a unified framework for joint segmentation and classification that incorporates physics-based augmentation and an efficiency-aware design. The approach models scanner-dependent variations using feature-wise statistics (mean and variance) via a frozen descriptor, enabling distribution alignment that normalizes hardware-induced differences while preserving anatomical structure. This improves robustness to domain shifts and reduces task interference.

Experiments on benchmark datasets show improved performance, achieving gains of 0.853 in segmentation and 0.918 in classification.

The main contributions are:

- A unified joint learning framework, PHYSJOINTNET, for simultaneous segmentation and classification with reduced task interference.
- A physics-based augmentation strategy using feature distribution alignment for robustness to domain shifts.
- Improved performance (2% classification and 4.1% segmentation) compared to existing approaches.

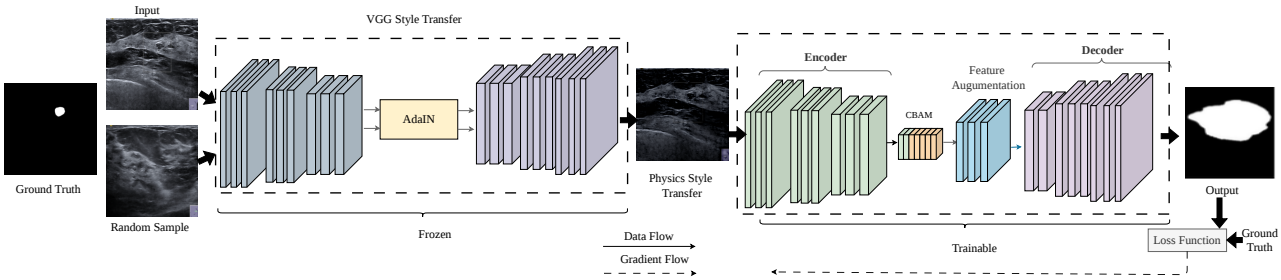

Figure 1. PHYJOINTNET framework. Input images are processed by a Frozen VGG Physics Descriptor to align hardware-specific texture statistics (μ, σ). These refined features undergo latent augmentation before task-specific heads predict the lesion mask and diagnostic label.

2. Problem Formulation

Let $\mathcal{D} = \{(x_i, y_i^c, y_i^s)\}_{i=1}^N$ denote a breast ultrasound dataset, where $x_i \in \mathbb{R}^{H\times W\times 3}$ is an input image, $y_i^c \in \{0, 1, 2\}$ represents diagnostic labels (*Normal*, *Benign*, *Malignant*), and $y_i^s \in \{0, 1\}^{H\times W}$ is the corresponding segmentation mask. The objective is to learn a function f_θ that jointly predicts classification and segmentation outputs: $f_\theta : x_i \mapsto (y_i^c, y_i^s)$.

This joint learning is challenging due to task interference, class imbalance in clinical data, and domain shifts caused by variations across ultrasound scanners. Additionally, deployment on point-of-care devices requires computational efficiency.

The learning objective is defined as:

$$\theta^* = \arg\min_\theta \mathbb{E}_{(x_i,y_i^c,y_i^s)\sim\mathcal{D}} \big[\mathcal{L}_{cls}(f_\theta^c(x_i), y_i^c) + \mathcal{L}_{seg}(f_\theta^s(x_i), y_i^s) \big], \quad (1)$$

where f_θ^c and f_θ^s denote classification and segmentation outputs, and $\mathcal{L}_{cls}$ and $\mathcal{L}_{seg}$ are the respective loss functions. The goal is to learn a robust and efficient model under these constraints.

3. Proposed Method

In this section, we describe the architecture of PHYJOINTNET (Fig. 1), a memory- and computation-efficient multi-task framework for robust breast ultrasound analysis. The framework consists of four key components: a physics-informed input augmentation module (Sec. 3.1), a quantized shared encoder (Sec. 3.2), an attention-driven feature refinement block (Sec. 3.3), and a dual-path latent augmentation strategy to mitigate class imbalance (Sec. 3.4). The network is trained end-to-end using a joint multi-task loss under strict low-bit precision constraints (Sec. 3.6).

3.1. Physics-Based Input Augmentation

Breast ultrasound images exhibit substantial variation in contrast, gain, and hardware-specific textures due to differences in device settings and acquisition environments. To simulate these variations while preserving anatomical structure, we introduce a Style-Stats Shift module based on a frozen VGG-19 physics descriptor. Let $x_i \in \mathbb{R}^{H\times W\times 3}$ denote a content image and $x_j \in \mathbb{R}^{H\times W\times 3}$ a reference style image sampled from a different scanner. We extract intermediate feature maps using the VGG encoder $f(\cdot)$ and perform channel-wise feature normalization using Adaptive Instance Normalization (AdaIN) [11]. The transformed input $\tilde{x}_i$ is computed as

$$\tilde{x}_i = \sigma\big(f(x_j)\big) \left(\frac{f(x_i) - \mu(f(x_i))}{\sigma(f(x_i))} \right) + \mu\big(f(x_j)\big), \quad (2)$$

where $\mu(\cdot)$ and $\sigma(\cdot)$ denote the channel-wise mean and standard deviation of the feature maps, computed across spatial dimensions. In Eq. (2), $\tilde{x}_i$ represents the augmented input that retains the anatomical content of x_i while adopting the scanner-specific style of x_j. During training, this transformation is applied with probability $p = 0.3$, exposing the network to a diverse set of acquisition variations. By normalizing hardware-dependent texture statistics in this manner, the model learns to focus on diagnostically relevant anatomical features, thereby improving generalization across heterogeneous scanners while remaining compatible with low-bit (INT4/INT8) deployment.

3.2. Shared Encoder

To enable deployment in resource-constrained clinical environments, the shared encoder $e(\cdot)$ is implemented using an EfficientNet-B6 backbone. The encoder weights θ_e are optimized to learn compact and expressive feature representations. Given an augmented input image $\tilde{x}_i \in \mathbb{R}^{H\times W\times 3}$ (from Eq. (2)), the encoder produces a latent feature tensor $z_i \in \mathbb{R}^{H'\times W'\times C}$ as

$$z_i = e(\tilde{x}_i; \theta_e), \quad (3)$$

where H' and W' are the spatial dimensions and C is the number of feature channels. This design enables extraction of rich multi-scale features $\mathcal{F}$ necessary for accurate lesion localization while maintaining efficiency for practical deployment.

3.3. Attention-Driven Feature Refinement

The latent features z_i are refined to suppress irrelevant background and enhance diagnostically relevant regions using the Convolutional Block Attention Module (CBAM) $\phi(\cdot)$. CBAM computes a channel attention map $M_c(z_i)$ to emphasize informative acoustic biomarkers and a spatial attention map $M_s(z_i')$ to localize lesion structures within the feature map. The refinement is applied sequentially as

$$z_i' = M_c(z_i) \otimes z_i, \quad z_i'' = M_s(z_i') \otimes z_i', \tag{4}$$

where $\otimes$ denotes element-wise multiplication, z_i' is the intermediate feature after channel attention, and z_i'' is the final refined feature. This attention-driven refinement (Eq. (4)) ensures that downstream task-specific heads focus on salient lesion cues while mitigating the effects of low-bit quantization.

3.4. Latent Feature Augmentation

To address class imbalance in the dataset, particularly for *Normal* and *Malignant* categories, we perform latent-level augmentation. For a minority-class sample, stochastic Gaussian noise $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ is injected into the refined features z_i'' to simulate intra-class echotexture variability. Additionally, intra-class feature interpolation generates synthetic latent features

$$\tilde{z}_{mix} = \lambda z_i + (1-\lambda) z_j, \quad \lambda \sim \text{Uniform}(0.3, 0.7), \tag{5}$$

where z_i and z_j belong to the same diagnostic category. The augmented features $\tilde{z}_i$ are then used as inputs to the multi-task heads.

3.5. Multi-Task Forks

The refined and augmented latent features $\tilde{z}_i$ are processed by two task-specific heads. The segmentation decoder $d_s(\cdot)$ employs a UNet++ style architecture with nested skip connections to produce pixel-wise lesion masks $\hat{y}_i^s \in [0,1]^{H \times W}$, preserving fine boundary details despite low-bit encoding. Simultaneously, the classification head $d_c(\cdot)$ applies global average pooling to z_i'' to obtain a compact feature vector v_i, which is passed through a linear-ReLU-softmax module to predict diagnostic probabilities $\hat{y}_i^c \in [0,1]^3$ for *Normal*, *Benign*, and *Malignant* classes.

3.6. Joint Optimization

The network parameters θ are optimized using a weighted joint loss function that combines segmentation and classification objectives:

$$\mathcal{L}(\theta) = \alpha\, \mathcal{L}_{DiceFocal}(y_i^s, \hat{y}_i^s) + (1-\alpha)\, \mathcal{L}_{WCE}(y_i^c, \hat{y}_i^c), \tag{6}$$

where y_i^s and y_i^c are the ground-truth segmentation mask and diagnostic label, $\hat{y}_i^s$ and $\hat{y}_i^c$ are the predictions, $\mathcal{L}_{DiceFocal}$ addresses pixel-level sparsity, $\mathcal{L}_{WCE}$ is the Weighted Cross-Entropy loss for class imbalance, and $\alpha \in [0,1]$ balances the tasks. Optimization is performed using 8-bit Adam (AdamW8bit) to complement the 4-bit encoder weights. The final objective is to find the optimal quantized parameters θ^* by minimizing the total joint loss over the dataset $\mathcal{D}$ of N samples:

$$\theta^* = \arg\min_{\theta \in \mathbb{INT}4} \sum_{i=1}^{N} \mathcal{L}(\theta). \tag{7}$$

Eqs. (3)–(7) define a fully quantized, multi-task learning framework capable of accurate lesion localization and classification under strict memory and computational constraints.

4. Experimental Details

4.1. Datasets

We evaluate the proposed framework on a subset of the PRECISE 2025 Challenge dataset [1, 6, 17], a multi-institutional breast ultrasound collection consisting of $3,120$ images with expert-annotated segmentation masks and diagnostic labels across three categories: *Normal*, *Benign*, and *Malignant*. Malignant cases comprise approximately 18% of the dataset, reflecting realistic clinical class imbalance. For additional validation, we also consider the Curated BUSI dataset [2], containing 450 images after duplicate removal, and BUS-UCLM [21] with 683 images, each exhibiting distinct class distributions. All datasets are partitioned in a patient-wise manner to prevent data leakage, ensuring that images from the same patient appear exclusively in a single split during cross-validation.

4.2. Preprocessing and Implementation Details

All images are resized to 256×256 using bilinear interpolation with zero-padding to preserve aspect ratio, and normalized using ImageNet statistics [5]. To improve generalization, data augmentation is applied during training, including random flips (p=0.5), rotation ($\pm 30°$), scaling (0.9–1.1), translation (up to 10%), brightness/contrast variations (±15%–20%), and Gaussian noise ($\sigma = 0.01$).

The framework is implemented in PyTorch 1.13.0 with CUDA 11.7 and trained on dual NVIDIA T4 GPUs (16 GB each). Convolutional layers use Kaiming initialization [8], while fully connected layers use Xavier initialization. Optimization is performed using AdamW [13] with a learning rate of 1×10^{-4}, weight decay of 1×10^{-2}, and batch size of 16. Training is conducted for 200 epochs using 4-fold cross-validation with early stopping (patience = 10) based on validation Dice score. Mixed-precision training (FP16) via `torch.cuda.amp` reduces memory usage by 40% without performance loss.

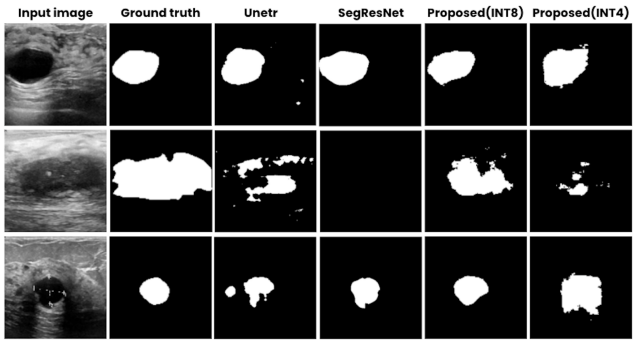

Figure 2. Qualitative segmentation results on BUS UCLM and Curated BUSI datasets.

Table 1. **Quantitative comparison** on the BUS-UCLM dataset.

Method	DSC ↑	JACC ↑	F1 ↑	AUC ↑
UNETR	0.734	0.579	0.619	0.790
SwinUNETR	0.805	0.673	0.254	0.500
nnU-Net	0.793	0.657	0.531	0.860
DeepLabV3+	0.805	0.796	0.857	0.967
Without PhyAug	0.812	0.748	0.899	**0.973**
Proposed	**0.853**	**0.788**	**0.918**	0.968

4.3. Comparison with State-of-the-Art

Qualitative Results: We compare our approach with UNETR [7], SegResNet [14], and precision variants on the BUSI dataset. The INT8 model produces sharper boundaries and better structural preservation compared to baselines. INT4 shows mild loss in fine detail and contour smoothness but remains competitive, benefiting from physics-based augmentation that improves robustness while lower precision limits spatial refinement.

Quantitative Results: We evaluate PhyJointNet on the BUS-UCLM dataset (Table 1). Baseline segmentation models report HD95 of 10.06–15.76 mm and DSC of 0.734–0.805. In contrast, PhyJointNet (FP32) improves DSC to 0.853 (+4.8%), and Jaccard to 0.788 (+17.0%). Even without physics augmentation, it outperforms baselines (DSC=0.812), showing strong backbone effectiveness. For classification, single-task segmentation models fail (AUC=0.500), while DeepLabV3+ achieves 0.967. PhyJointNet (FP32) reaches 0.968 AUC and 0.918 F1, slightly surpassing the best baseline, confirming that joint learning improves performance without harming segmentation.

4.4. Ablation Study

Physics-informed Augmentation. Fig. 3 shows the impact of integrating physics-informed style augmentation with the baseline model. Among the evaluated strategies, the VGG Style-Stats Shift module provides the largest gain, improving segmentation DSC from 0.812 to 0.853 and classification F1 from 0.899 to 0.918. Qualitative results further show more stable predictions across different scanners, with better boundary delineation and reduced sensitivity to intensity variations. These improvements demonstrate that modeling acquisition-related variability is beneficial. By simulating physics-driven changes such as gain, contrast, and intensity shifts, the augmentation encourages the model to focus on anatomical structures rather than scanner-specific artifacts, improving generalization across heterogeneous clinical settings.

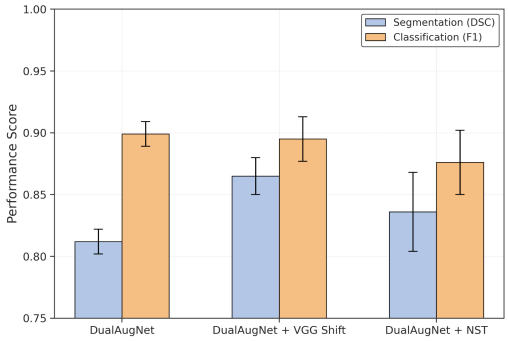

Figure 3. Impact of physics-informed style augmentation on multi-task performance.

Memory Efficiency and Deployment. The computational and storage requirements of the proposed model indicate that the number of floating-point operations (4.56 G FLOPs) and model parameters (41.48 M) remain fixed, reflecting a stable architectural complexity. While the model maintains high representational capacity, its design supports efficient deployment on resource-constrained hardware platforms such as edge devices and embedded medical systems, where storage and bandwidth are critical. Furthermore, the architecture facilitates faster data transfer, improved cache utilization, and reduced energy consumption during inference, making it well-suited for real-time clinical applications without requiring modifications to the core network.

5. Conclusion

We proposed PHYJOINTNET, a unified framework for joint segmentation and classification in breast ultrasound. The model incorporates physics-informed augmentation to improve robustness against scanner-induced domain shifts and reduce task interference in multi-task learning. Experiments show that PHYJOINTNET outperforms existing baselines, with gains of 4.1% in segmentation and 2% in classification. Future work will focus on real-time deployment on edge hardware such as FPGA/ASIC platforms for low-latency clinical use.

References

[1] Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in brief*, 28:104863, 2020. 3

[2] Carlos Aumente-Maestro, Jorge Díez, and Beatriz Remeseiro. A multi-task framework for breast cancer segmentation and classification in ultrasound imaging. *Computer methods and programs in biomedicine*, 260:108540, 2025. 3

[3] Wendie A Berg, Jeffrey D Blume, Jean B Cormack, Ellen B Mendelson, Daniel Lehrer, Marcela Böhm-Vélez, Etta D Pisano, Roberta A Jong, W Phil Evans, Marilyn J Morton, et al. Combined screening with ultrasound and mammography vs mammography alone in women at elevated risk of breast cancer. *Jama*, 299(18):2151–2163, 2008. 1

[4] Michael Crawshaw. Multi-task learning with deep neural networks: A survey. *arXiv preprint arXiv:2009.09796*, 2020. 1

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009. 3

[6] Wilfrido Gómez-Flores, Maria Julia Gregorio-Calas, and Wagner Coelho de Albuquerque Pereira. Bus-bra: a breast ultrasound dataset for assessing computer-aided diagnosis systems. *Medical Physics*, 51(4):3110–3123, 2024. 3

[7] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI brainlesion workshop*, pages 272–284. Springer, 2021. 4

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034, 2015. 3

[9] Qinghua Huang, Yaozhong Luo, and Qiangzhi Zhang. Breast ultrasound image segmentation: a survey. *International journal of computer assisted radiology and surgery*, 12(3):493–507, 2017. 1

[10] Qing Huang et al. Joint detection and classification of breast lesions in ultrasound images using deep learning. *Medical Physics*, 2020. 1

[11] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, 2017. 1, 2

[12] Taminul Islam, Md Alif Sheakh, Mst Sazia Tahosin, Most Hasna Hena, Shopnil Akash, Yousef A Bin Jardan, Gezahign FentahunWondmie, Hiba-Allah Nafidi, and Mohammed Bourhia. Predictive modeling for breast cancer classification in the context of bangladeshi patients by use of machine learning approach with explainable ai. *Scientific Reports*, 14(1):8487, 2024. 1

[13] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 3

[14] Andriy Myronenko. 3d mri brain tumor segmentation using autoencoder regularization. In *International MICCAI brainlesion workshop*, pages 311–320. Springer, 2018. 4

[15] J Alison Noble and Djamal Boukerroui. Ultrasound image segmentation: a survey. *IEEE Transactions on medical imaging*, 25(8):987–1010, 2006. 1

[16] Bassamma Patil, P Vishwanath, K Priyanka, Muhamed Husseyn, and KG Parthiban. Convolutional neural network–regularized extreme learning machine with hyperbolic secant for breast cancer segmentation and classification. In *2025 3rd International Conference on Integrated Circuits and Communication Systems (ICICACS)*, pages 1–6. IEEE, 2025. 1

[17] Anna Pawłowska, Anna Ćwierz-Pieńkowska, Agnieszka Domalik, Dominika Jaguś, Piotr Kasprzak, Rafał Matkowski, Łukasz Fura, Andrzej Nowicki, and Norbert Żołek. Curated benchmark dataset for ultrasound based breast lesion analysis. *Scientific Data*, 11(1):148, 2024. 3

[18] Khosro Rezaee, Adele Rezaee, Negar Shaikhi, and Javad Haddadnia. Multi-mass breast cancer classification based on hybrid descriptors and memetic meta-heuristic learning. *SN Applied Sciences*, 2(7):1297, 2020. 1

[19] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 1

[20] Swati Shilaskar, Shripad Bhatlawande, Mayur Talewar, Sidhesh Goud, Soham Tak, and Sachi Kurian. Classification and segmentation of breast tumor ultrasound images using vgg-16 and unet. *Biomedical and Pharmacology Journal*, 18(1): 569–580, 2025. 1

[21] Noelia Vallez, Gloria Bueno, Oscar Deniz, Miguel Angel Rienda, and Carlos Pastor. Bus-uclm: Breast ultrasound lesion segmentation dataset. *Scientific Data*, 12(1):242, 2025. 3

[22] Xin Yi, Ekta Walia, and Paul Babyn. Generative adversarial network in medical imaging: A review. *Medical image analysis*, 58:101552, 2019. 1