Technical Article

DAUS-Net: Toward Ultrasound Scanner-Agnostic Domain Generalized Robust and Accurate Segmentation

Ultrasonic Imaging
2026, Vol. 48(3) 201–214

DOI: 10.1177/01617346251388454
journals.sagepub.com/home/uix
Mary Ann Liebert
A Part of Sage

Sangheon Lee[1], Dongkyu Jung[1], Nizar Guezzi[1], Sangwoo Nam[1], and Jaesok Yu[1,2,3]

Abstract

In medical imaging, segmentation is a critical task for analysis and diagnosis. Deep learning-based segmentation has been actively studied and has shown remarkable performance. Building high-accuracy segmentation models requires a large amount of high-quality labeled data, but the cost of collecting such data is extremely high in medical imaging. In ultrasound imaging, the differences in image features depending on the equipment are significantly greater compared to other medical imaging modalities. Consequently, models need to be trained for each specific device, which entails substantial costs and time, leading to various practical challenges. To address these challenges, we propose a robust and accurate segmentation network that can operate independently of the ultrasound equipment. We integrated the Deep Frequency Filtering (DFF) module into a U-Net-based model. The proposed model retains the U-Net's encoder-decoder structure but applies frequency filtering within the latent space of each encoder layer, enabling adaptive selection of frequency components for breast tumor detection. Moreover, batch normalization was replaced with instance normalization to remove stylish features. We evaluated the model using three public datasets acquired from different scanners, achieving superior performance on unseen testing datasets compared to existing models. Notably, when tested on the unseen BUS-BRA dataset, DAUS-Net achieved a Dice score of 0.76, compared to 0.61 by the conventional U-Net. This improvement is attributed to the synergy between the DFF module and instance normalization. Our results demonstrate that the proposed model consistently detects and segments breast tumors, highlighting its potential for generalized clinical segmentation task. The source code for implementing DAUS-Net is publicly available at https://github.com/shlee8638/DAUS-Net.

Keywords

deep frequency filtering, deep learning, domain generalization, scanner-agnostic learning, ultrasound image segmentation

Introduction

Ultrasound segmentation, particularly for detecting and diagnosing tumors, cysts, and other anomalies, has emerged as a crucial task in various clinical applications, particularly in obstetrics, cardiology, and oncology.[1-3] Accurate segmentation of ultrasound images is essential for precise measurements, monitoring disease progression, and planning treatments, making it indispensable for early diagnosis and intervention.[4] Segmenting ultrasound images remains a challenging task due to several factors unique to this imaging modality. One major challenge lies in the inherent noise and speckle artifacts present in ultrasound images, which complicate the process of identifying clear boundaries between tissues.[1-5] Moreover, variations in equipment, patient anatomy, and even operator skills introduce additional complexity to segmentation, often leading to inconsistent results.[1,6,7] As a result, the task of segmenting objects like tumors, organs, and lesions from ultrasound images is particularly difficult, requiring advanced algorithms to overcome these challenges.

Deep learning techniques, such as U-Net and several modified networks, have shown great promise in automating the segmentation of ultrasound images. Building on this progress, recent directions include saliency-guided attention,[8] detect-then-segment pipelines with test-time augmentation,[9]

[1]Department of Robotics and Mechatronics Engineering, DGIST, Daegu, Republic of Korea
[2]Department of Biomedical Science and Engineering, DGIST, Daegu, Republic of Korea
[3]Department of the Interdisciplinary Studies of Artificial Intelligence, DGIST, Daegu, Republic of Korea

Corresponding Author:
Jaesok Yu, Department of Robotics & Mechatronics Engineering, Department of Biomedical Science and Engineering, Department of the Interdisciplinary Studies of Artificial Intelligence, Daegu Gyeongbuk Institute of Science and Technology (DGIST), Daegu 42988, Republic of Korea.
Email: jaesok.yu@dgist.ac.kr

adaptive receptive fields via selective kernels fusing dilated and standard convolutions,[10] residual and extended-residual encoders with mixed attention loss,[11] cascaded global guidance with residual refinement,[12] foreground–background saliency with morphology-aware fusion,[13] hybrid channel and spatial self-attention in place of convolutions,[14] shared-weight nested depths,[15] global feature mixing with attention-gated skips,[16] and enhanced selective kernels with deep supervision.[17] These models demonstrate strong performance in extracting meaningful features and distinguishing tissue types. However, deployment in real clinical environments remains difficult. Devices vary across manufacturers and scanner models, and beamforming and receive chains change the point spread function and noise. Transducer type and center frequency alter resolution, attenuation, and speckle grain. Operator settings such as gain, time gain compensation, dynamic range, and log compression reshape brightness and contrast by depth. Probe handling and insonation angle influence shadowing and posterior enhancement. These differences shift texture, boundaries, and intensity statistics, which makes generalization across devices, sites, and patient populations challenging for segmentation models.[18] Furthermore, the lack of sufficiently labeled data and the need for high-quality training sets remain significant barriers to developing robust models that can perform well in diverse clinical scenarios.

Previous studies have reported that deep learning models, when trained on data collected in specific clinical environments, can achieve segmentation performance comparable to that of experienced clinicians. However, this performance is typically observed only when tested on data from the same domain. In real clinical scenarios, it is common for new devices to be introduced or existing equipment to be updated. In such cases, while experienced clinicians can achieve consistent outcomes using their anatomical knowledge and expertise, deep learning models with restricted generalization capabilities may exhibit a noticeable reduction in performance when applied to new environments.[19,20]

To address this issue, one potential solution is to collect data from the new devices and use it to fine-tune the mode.[19] However, the fine-tuning process is often impractical due to the high costs and time-consuming nature of labeling additional medical data, which requires skilled professionals to manually annotate it.[21] Consequently, retraining the model for every new device or domain shift becomes inefficient and resource-intensive. Instead, in the absence of input or label data from the unseen target domain, an alternative approach, known as Domain Generalization (DG), has been proposed to achieve high performance across diverse domains.[22]

Domain shifts in ultrasound imaging stem from various factors related to the data acquisition process. For instance, even slight variations in parameter settings during image acquisition can significantly affect the signal-to-noise ratio (SNR), and differences between manufacturers can result in substantial variations in image contrast.[18] These factors pose significant challenges for deep learning models in generalizing across diverse ultrasound imaging settings.

Another perspective to consider is the inherent challenges associated with the breast tumor ultrasound data used in this study. As shown in Figure 1, breast tumors vary widely in size and shape (e.g., benign vs. malignant) depending on the patient's condition. Additionally, the tumor boundaries are often ambiguous, which can increase label uncertainty and result in inter- and intra-rater variability among experts.[23-26] Moreover, the unique speckle noise in ultrasound imaging contributes to a lower signal-to-noise ratio (SNR) compared to other imaging modalities. Addressing these complex problems simultaneously is not straightforward. To tackle them, this study proposes a feature-based approach specifically designed for ultrasound imaging.

One approach to DG involves separating stylish features from anatomical features.[27,28] Stylish features originate from domain-specific characteristics, while anatomical features are derived from the underlying anatomical structure. Therefore, an ideal model should eliminate stylish features and rely exclusively on anatomical features to detect and segment the target. Studies employing this approach have often relied on additional learning paradigms. However, the most practical and feasible approach in clinical settings is to achieve robust generalization through a simplified learning process, without the need for additional learning paradigms.

Instance normalization (IN) is one such method.[29,30] From the perspective of DG, IN normalizes each image based on its statistics, removing its inherent style. However, IN not only removes stylish features but may also inadvertently eliminate discriminative information—defined in this study as the information necessary to distinguish between the breast tumor and the background. To overcome this limitation, new approaches are required.[31,32]

Meanwhile, the Deep Frequency Filtering (DFF) module introduced in[33] transforms intermediate features into the Fourier domain, builds a compact latent representation with 1×1 convolutions, and computes a spatial-attention mask in Fourier space that is applied before the inverse FFT. This mask functions as a content- and layer-adaptive frequency filter. From a physics standpoint, ultrasound exhibits frequency-dependent attenuation $\alpha(f) = \alpha_0 |f|^y$ and scanner-specific bandwidth/PSF differences, indicating that cross-center differences are largely spectral. From a signal perspective, speckle behaves multiplicatively and degrades contrast, whereas boundary cues concentrate at higher spatial frequencies. As visualized in Figure 9, the learned DFF mask, computed in the Fourier coordinate system and applied before the inverse FFT, selectively reweights spectral bands, preserving tumor-relevant anatomy while attenuating scanner-specific nuisance spectra in a content- and stage-adaptive manner (encoder depth). Additionally, while[33] incorporates a gradient reversal layer during training, it can be inferred that this was intended to guide the DFF module to filter out

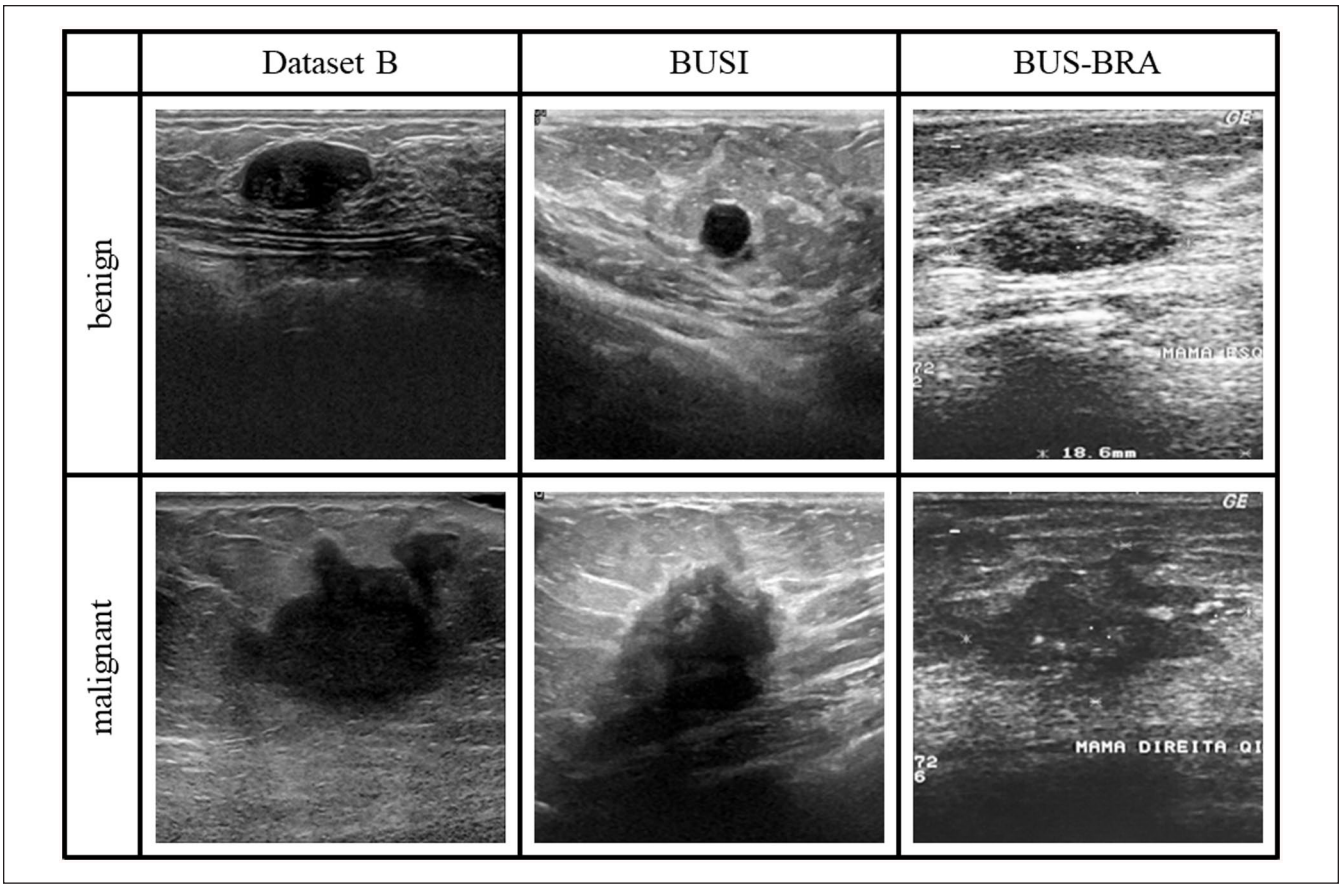

Figure 1. Examples of benign and malignant cases from the three public datasets used in the experiments.
Image courtesy of Refs.[41-43]

high-frequency components (HFC). This design choice is based on the observation that models optimized using stochastic gradient descent tend to prioritize learning low-frequency components (LFC) during the initial stages of training but gradually shift toward learning HFC to improve accuracy.[34,35] However, this shift comes at the cost of reduced robustness, which is traded off against accuracy, ultimately diminishing DG performance. Although the gradient reversal layer helps prevent the model from learning domain-specific features, it faces limitations in medical imaging, where dataset domains often overlap, making its application more challenging.

Therefore, this study proposes a novel approach that combines IN and the DFF module, leveraging their strengths while addressing their limitations to achieve generalized segmentation performance for ultrasound images, regardless of the device type, similar to the capabilities of a clinical expert. The primary limitation of IN is its tendency to remove discriminative information. However, the DFF module performs instance-adaptive frequency filtering with the breast tumor as the target, preserving meaningful frequency information specific to each image. Conversely, the DFF module often requires additional learning paradigms to facilitate DG. To overcome this, IN was employed to remove stylish features, mitigating domain differences and enhancing generalization performance without the need for additional learning paradigms.

Therefore, the contributions of this study are as follows:

1. We propose DAUS-Net, a novel ultrasound segmentation architecture that integrates Instance Normalization and Deep Frequency Filtering to achieve domain generalization without requiring adversarial training or access to target domain data.
2. The model performs instance-adaptive frequency filtering in the latent space, enabling robust extraction of anatomical features while suppressing domain-specific style variations.
3. We experimentally validate DAUS-Net under domain generalization settings using multiple ultrasound datasets acquired from diverse scanners, demonstrating consistent improvements over U-Net and SK-U-Net baselines on unseen domains.

As a result, this study enables robust and generalized detection and segmentation of breast ultrasound images, ensuring consistency across images despite domain shifts, such as device-specific variations in image characteristics.

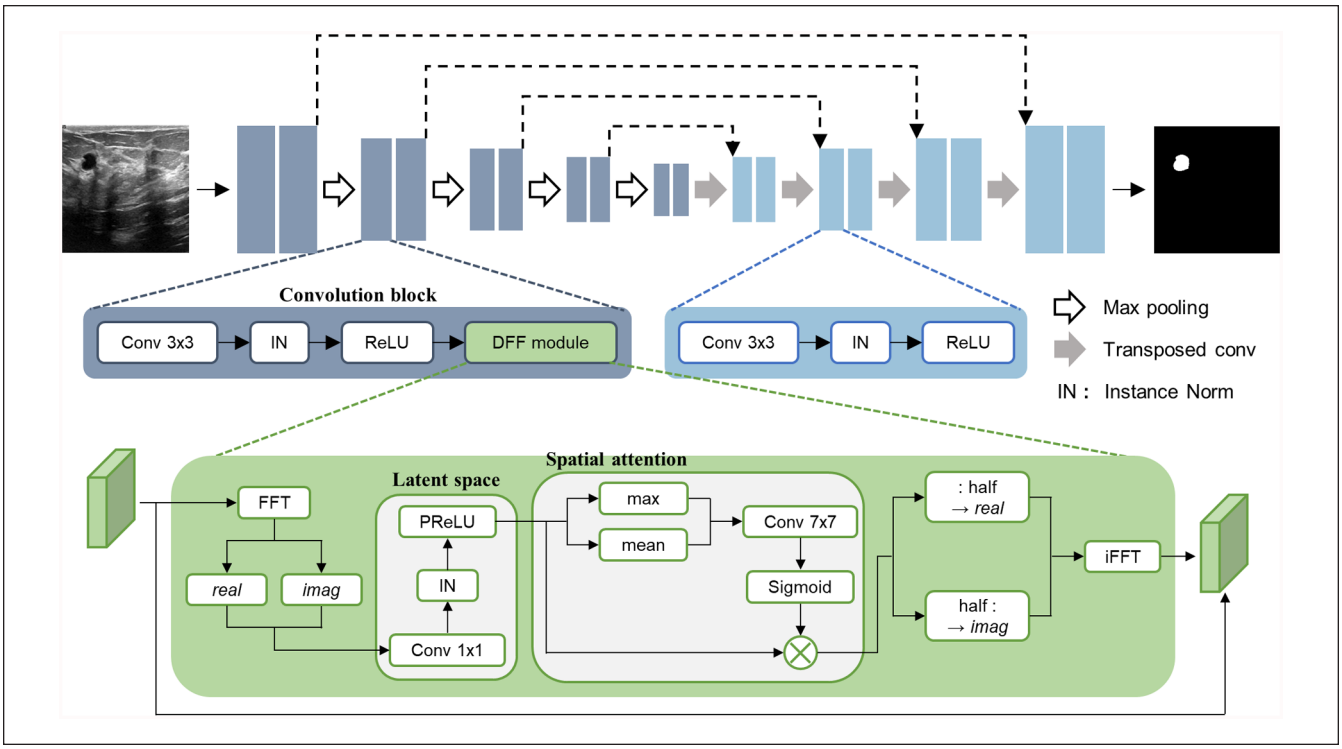

Figure 2. Overall architecture of DAUS-Net: Both the encoder and decoder follow a convolution block, but the encoder includes the DFF module. The encoder reduces the feature map size through max pooling, while the decoder restores the size through transposed convolution.

Methods

As shown in Figure 2, the proposed model, DAUS-Net (Domain-Agnostic Ultrasound Segmentation Network), adopts a U-Net-like architecture with an encoder-decoder structure comprising four down-sampling and four up-sampling stages.[36] However, it differs from the standard U-Net in two key aspects. First, it replaces of Batch Normalization (BN) with IN. Second, it incorporates the DFF module into the convolution blocks of the encoder. These modifications enable the model to generalize across different domains by leveraging frequency-based feature extraction and adapting to variations in image characteristics.

Deep Frequency Filtering Module

The DFF module was first introduced in the previous study.[33] Unlike conventional spatial attention, the DFF module operates in the frequency domain. The input feature map, $\mathrm{X} \in \mathbb{R}^{C\times H\times W}$, enters the DFF module, where it undergoes a Fourier transform for processing.

$$\mathrm{X}_1 = \mathcal{F}_{2D}\{\mathrm{X}\} \tag{1}$$

This transform provides global context at low cost and separates phase from magnitude, enabling boundary-sensitive phase information to be maintained while adaptively regulating frequency components to mitigate scanner-dependent low-frequency appearance drift. Due to Hermitian symmetry, its size can be reduced by half while preserving all necessary information. Consequently, the halved output, $X_1 \in \mathbb{C}^{C\times H\times\left(\left\lfloor\frac{w}{2}\right\rfloor+1\right)}$, is split into its real part and imaginary parts. These values are separated into channels and concatenated into a single tensor, resulting in $X_1 \in \mathbb{R}^{2C\times H\times\left(\left\lfloor\frac{w}{2}\right\rfloor+1\right)}$. The 1×1 layer linearly recombines the doubled real–imag channels to align phase relations and reduce cross-channel redundancy, producing a compact latent that is easier for attention to use. Finally, IN and PReLU are applied:

$$X_2 = PReLU(IN(Conv_{1\times1}(X_1))) \tag{2}$$

In the previous DFF module, BN was utilized[33]; however, in our proposed approach, we replaced BN with IN. IN mitigates scanner and preset appearance variation across centers, and PReLU preserves informative negative responses carried by phase sensitive channels. Following this, spatial attention is performed in the latent space.[37] Specifically, max-pooling and average-pooling operations are applied in the latent space, and the resulting features are concatenated to form the attention features.

$$X_3 = \text{concatenate}(X_2^{avg}, X_2^{max}). \quad (3)$$

In the spatial attention, a mask is generated using a 7×7 convolution, dynamically adapting to the input image. The mask is then refined with a sigmoid function to constrain its values between 0 and 1, facilitating the filtering process. The mask is conditioned on each location's spectral signature, which already encodes global context. It adaptively regulates spectral components, preserving boundary-sensitive phase cues while reducing sensitivity to low-frequency variations introduced by factors such as TGC or log compression.

$$\mathrm{M} = \text{Sigmoid}(Conv_{7\times7}(X_3)) \quad (4)$$

The mask M is then applied to X_2 through element-wise multiplication, serving as a content- and stage-adaptive frequency filter.

$$X_4 = X_2 \otimes \mathrm{M} \quad (5)$$

The resulting feature after spatial attention contains $2\mathrm{C}$ channels. These are divided into two-halves, representing the real and imaginary parts, which are then recombined into a complex tensor, $X_5 \in \mathbb{C}^{C\times H\times\left(\left\lfloor\frac{w}{2}\right\rfloor+1\right)}$. Finally, X_5 undergoes an inverse Fourier transform to produce an output with the same dimensions as the input.

$$X_6 = \mathcal{F}_{2D}^{-1}\{X_5\} \quad (6)$$

In the final step, the input and output tensors are combined through element-wise summation.

Instance Normalization

BN normalizes each channel across a mini-batch.[38] During the training phase, BN learns the parameters required for normalization from the training dataset. However, in the test phase, these learned parameters are fixed and applied to generate outputs. If the testing dataset exhibits domain characteristics that differ from those of the training dataset, this can result in degraded model performance. In contrast, IN normalizes each channel independently for a single image.[39] In this experiment, IN was configured without any learnable parameters, ensuring normalization to a mean of 0 and a variance of 1 for each channel within an image. This type of normalization effectively removes domain-specific styles. IN is defined as follows:

$$\mathrm{IN}(x) = \frac{x - \mu_i}{\sqrt{\sigma_i^2 + \varepsilon}} \quad (7)$$

μ_i and σ_i denote the mean and standard deviation of each channel within a single image.

Loss Function

For the loss function, we used binary cross-entropy, which is commonly employed in general segmentation task.[40] For breast images, binary cross-entropy is particularly suitable as the task focuses on separating the tumor from the background.

$$\begin{aligned}\mathcal{L}_{BCE} = &-\sum_{i,j} y(i,j)\cdot\log\hat{y}(i,j) \\ &+\left(1-y(i,j)\right)\cdot\log(1-\hat{y}(i,j))\end{aligned} \quad (8)$$

In this equation, $\mathrm{y(i,j)} \in \{0,1\}$ denotes the ground truth for each pixel, while $\hat{y}(i,j) \in [0,1]$ represents the corresponding pixel value in the predicted mask.

Datasets and Experimental Setting

Datasets

To evaluate the robustness of DG, we utilized five public datasets in our experiments. The first dataset, Dataset B, was collected at the UDIAT Diagnostic Center of the Parc Taulí Corporation in Sabadell, Spain, using a Siemens ACUSON Sequoia C512 system with a 17L5 HD linear array transducer (8.5 MHz).[41] It includes 163 images with a mean resolution of 760 pixels × 570 pixels, comprising 53 malignant and 110 benign lesions. The second dataset, BUSI, was collected in 2018 at Baheya Hospital using a LOGIQ E9 and a LOGIQ E9 Agile ultrasound system with transducers operating at 1 to 5 MHz on an ML6-15-D Matrix linear probe.[42] The dataset consists of 780 images with a mean resolution of 500 pixels × 500 pixels, including 210 malignant, 487 benign, and 133 normal cases. For this study, normal cases were excluded. The third dataset, BUS-BRA, was collected at the National Institute of Cancer in Rio de Janeiro, Brazil, including the GE Logiq 5 (10-12 MHz), GE Logiq 7 (10-14 MHz), Toshiba Aplio 300 (12-14 MHz), and GE U-Systems (12-14 MHz).[43] It comprises 1875 images, with a mean resolution of approximately 320 × 400 pixels, including 607 malignant and 1268 benign lesions. The fourth dataset, BUS-UCLM, comprises 683 images from 38 patients acquired at Ciudad Real General University Hospital using a Siemens ACUSON S2000 system with an 18L6 HD linear probe.[44] All images have fixed dimensions of 768 pixels × 1024 pixels. The dataset includes 419 normal, 174 benign, and 90 malignant cases. For this study, normal cases were excluded. The fifth dataset, **BLUI**, was acquired at Shahid Beheshti University of Medical Sciences using a SuperSonic Imagine system with a 5 to 18 MHz linear transducer.[45] The dataset comprises **232 images** with radiologist-provided annotations, including **109 benign** and **123 malignant** lesions, each confirmed by histopathology (Table 1).

Table 1. Summary of Datasets.

Dataset name	# cases with masses	# Benign findings	# Malignant findings	# Unique patients	Scanner and acquisition settings	Image size (px)
Dataset B	163	110	53	163	Siemens ACUSON Sequoia C512, 17L5 HD linear array, 8.5 MHz	760 × 570
BUSI	697	487	210	600	LOGIQ E9/Agile, ML6-15-D Matrix linear probe, 1 to 5 MHz	500 × 500
BUS-BRA	1875	1268	607	1064	GE Logiq 5/7, Toshiba Aplio 300, GE U-system, linear array, 10-14 MHz	300 × 400
BUS-UCLM	264	174	90	38	Siemens Acuson S2000, 18L6 HD probe	768 × 1024
BLUI	232	109	123	Not reported	AixPlorer Ultimate ultrasound machine, linear transducer, 5 to 18 MHz	570 × 500

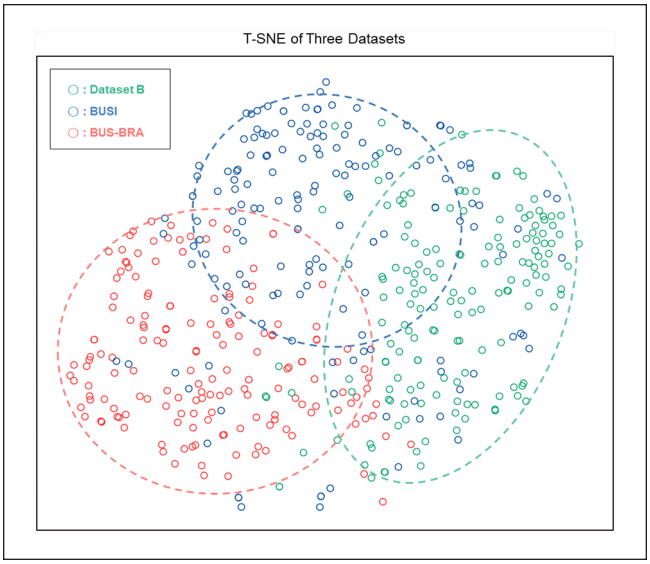

Figure 3. T-SNE visualization of three datasets: Dataset B, BUSI, BUS-BRA. The visualization shows that while there are overlapping regions in the distributions of the three datasets, each dataset fundamentally exhibits its own unique distribution. In the case of BUS-BRA, the dataset was collected using multiple transducers, and for the purpose of T-SNE analysis, data with lower SNR was primarily selected.

As shown in Figure 3, T-SNE visualization was conducted by sampling images to analyze the distribution of the three datasets.[46] For the BUS-BRA dataset, samples with lower SNR were prioritized. The red circles indicate BUS-BRA, the blue circles represent BUSI, and the green circles correspond to Dataset B. The T-SNE visualization reveals that, unlike general images, the distributions of the three datasets partially overlap while maintaining their distinct distributions.

Experimental Settings

In our experimental design, we conducted two complementary protocols. In the first, Dataset B and BUSI were used for training and validation, while the remaining datasets (BUS-BRA, BUS-UCLM, and BLUI) were reserved for testing. In the second, Dataset B, BUS-BRA, BUS-UCLM, and BLUI were used for training and validation, and BUSI was held out for testing. To construct the training and validation sets, we applied fivefold cross-validation while preserving the ratio of benign and malignant cases. For each protocol, the corresponding test set comprised the entire dataset of the held-out domain. For the BUS-BRA dataset, the provided fivefold cross-validation splits were used. When combining multiple datasets for training and validation, we merged them before applying the split procedure. All input images were resized to 256 × 256, and data augmentation was applied.

The Adam optimizer was employed with an initial learning rate of 0.001, which was reduced by a factor of 0.98 if the validation loss did not improve for three consecutive epochs.[44] Each experiment ran for 100 epochs with a batch size of 16, and the model with the lowest validation loss was saved for testing. All experiments were performed on a single NVIDIA RTX A5000 GPU, implemented using Python 3.10.16 and PyTorch 2.5.0.[47]

For comparison, we selected U-Net[36] and its variant SK-U-Net[10] as baselines, as they are widely used and well-established in Breast Ultrasound Segmentation (BUS) tasks. In addition, we included NU-Net,[15] CMU-Net,[16] and ESKNet,[17] which represent recent architectures specifically designed for BUS and have demonstrated competitive performance. These models also share similar architectural complexity with our proposed method, enabling a fair and comprehensive evaluation. We did not include Vision Transformer-based segmentation models due to their significantly larger parameter sizes and data requirements. Such models typically require large-scale datasets to achieve stable training and generalization, making them less suitable for relatively small ultrasound datasets like those used in this study.

Evaluation Metrics

The evaluation metrics included accuracy, Dice score, Intersection over Union (IoU), sensitivity, specificity, and the Area Under the ROC Curve (AUC).[48] To assess statistical significance on a per-image basis, we compared DAUS-Net

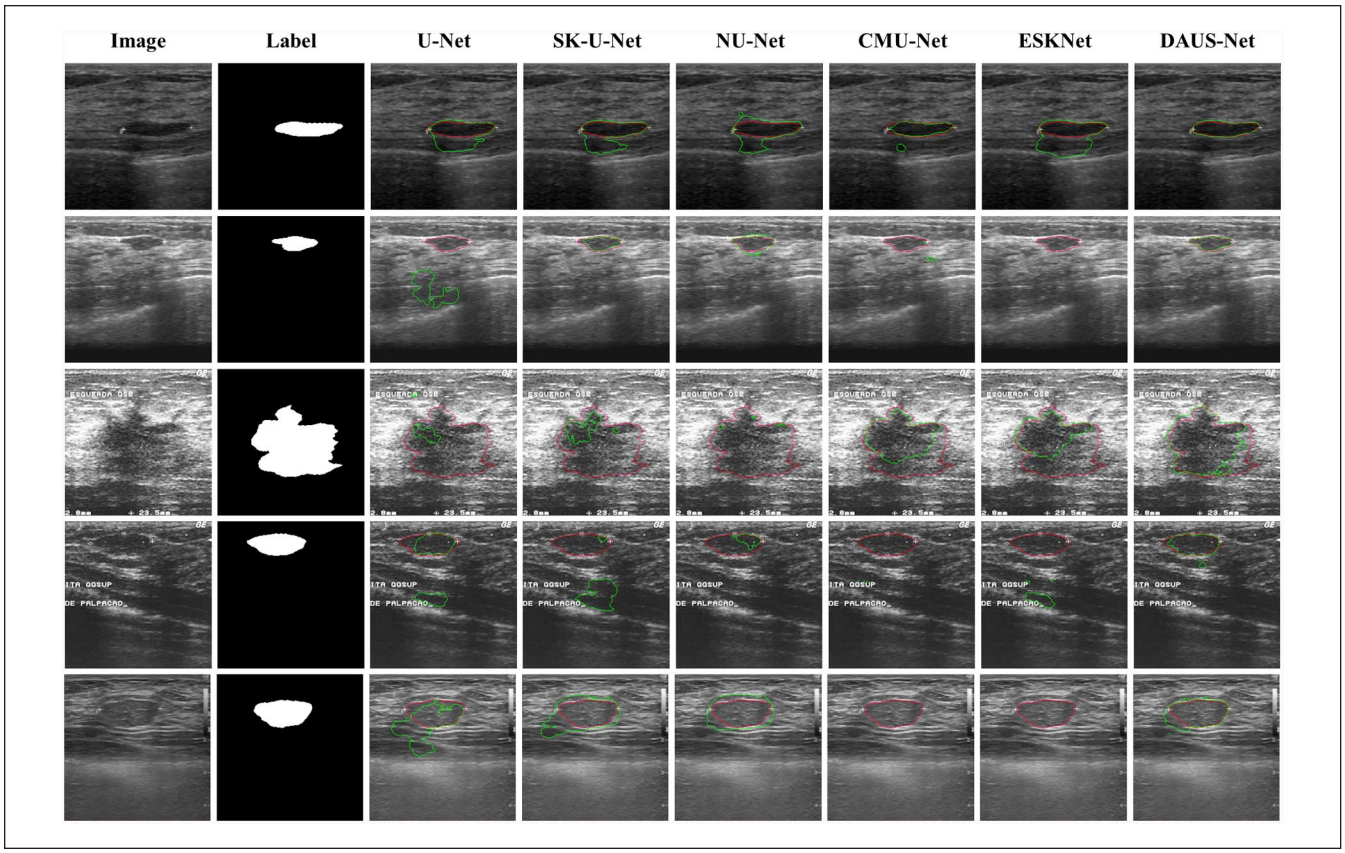

Figure 4. Predicted masks of U-Net, SK-U-Net, NU-Net, CMU-Net, ESKNet, and DAUS-Net trained/validated on Dataset B and BUSI, and tested on BUS-BRA. The red boundaries represent the label, and the green boundaries represent the predicted mask.

with each baseline using two-sided Wilcoxon signed-rank tests across all metrics.[49] Multiplicity across metrics and baselines was controlled with the Benjamini–Hochberg false discovery rate, and we report q-values with a threshold of 0.05.[50] Practical significance was quantified using Cohen's d for paired samples.[51] In the results tables, † denotes $q < 0.05$ and ↑ denotes $d_z \geq 0.5$.

Experimental Results

Figures 4 to 7 presents the predicted masks generated by the proposed model and comparison models across the four experiments. For clarity, the training/validation and test datasets employed in each experiment are explicitly indicated in the figure captions and in the titles of Tables 2 to 5. Green boundaries indicate the predicted mask, while red boundaries represent the ground truth. These boundaries are overlaid on the input images to highlight the regions the model focused on during mask prediction.

In some datasets, lesion boundaries are relatively well-defined, whereas in others, the boundaries appear indistinct with lower contrast, leading to reduced performance of the comparison models. Despite this, the proposed model, while not achieving perfect segmentation, consistently detects tumor regions with high accuracy. This observation will be further analyzed in the Discussion section.

Tables 2 to 5 present the mean evaluation metrics for the three experiments. Each table reports Dice, IoU, Accuracy (Acc), Sensitivity (Sens), Specificity (Spec), and AUC. Values are presented as mean ± standard deviation over five-fold cross-validation. Symbols are used to indicate statistical significance: † denotes $q < 0.05$, and ↑ denotes $d_z \geq 0.5$.

The results indicate that the performance of the comparison models varies across experimental settings, with certain tables showing relatively high Dice scores and IoU values and others exhibiting generally lower values. In Table 4, where the overall performance is comparatively high, the gap between the proposed model and the baselines is narrower, yet the proposed model consistently achieves the best mean performance. Conversely, in Table 3, where all models perform worse overall, the proposed model still attains the highest Dice score and IoU. In Table 2, comparisons between the proposed model and the baselines reveal that the majority of metrics yield q values below 0.05. Moreover, for most baselines, the effect size d_z exceeds 0.5 in Dice, IoU, and Sens, indicating both statistical and practical significance.

Table 4 and Figure 7 present the results obtained by training and validating on four datasets (Dataset B, BUS-BRA,

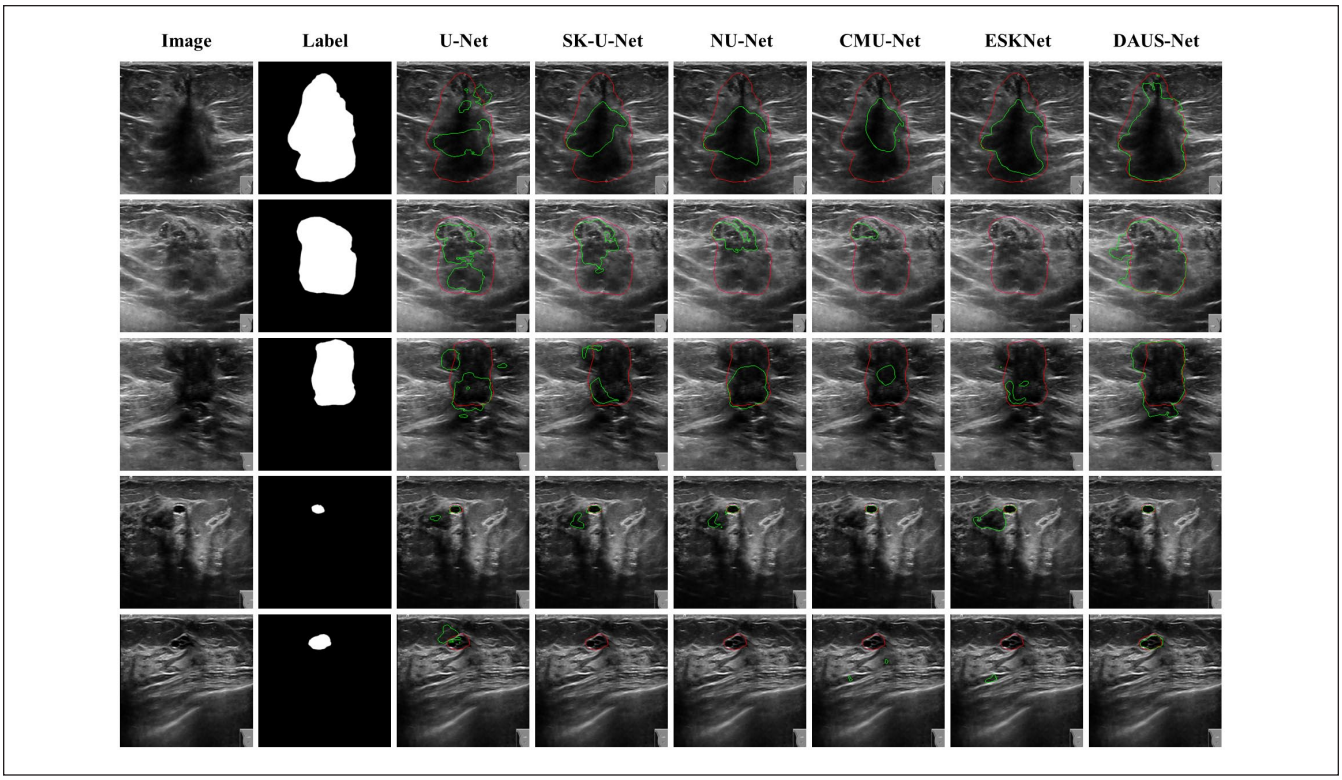

Figure 5. Predicted masks of U-Net, SK-U-Net, NU-Net, CMU-Net, ESKNet, and DAUS-Net trained/validated on Dataset B and BUSI, and tested on BUS-UCLM. The red boundaries represent the label, and the green boundaries represent the predicted mask.

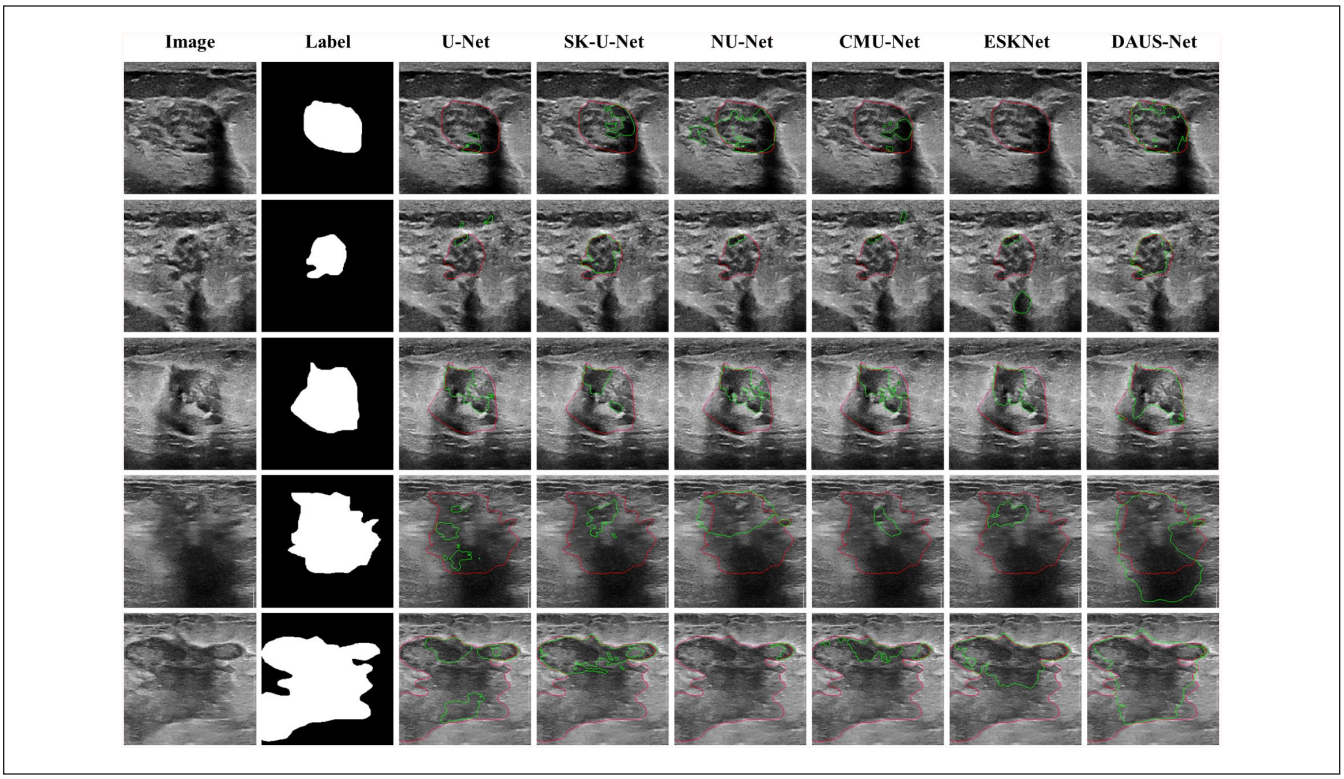

Figure 6. Predicted masks of U-Net, SK-U-Net, NU-Net, CMU-Net, ESKNet, and DAUS-Net trained/validated on Dataset B and BUSI, and tested on BLUI. The red boundaries represent the label, and the green boundaries represent the predicted mask.

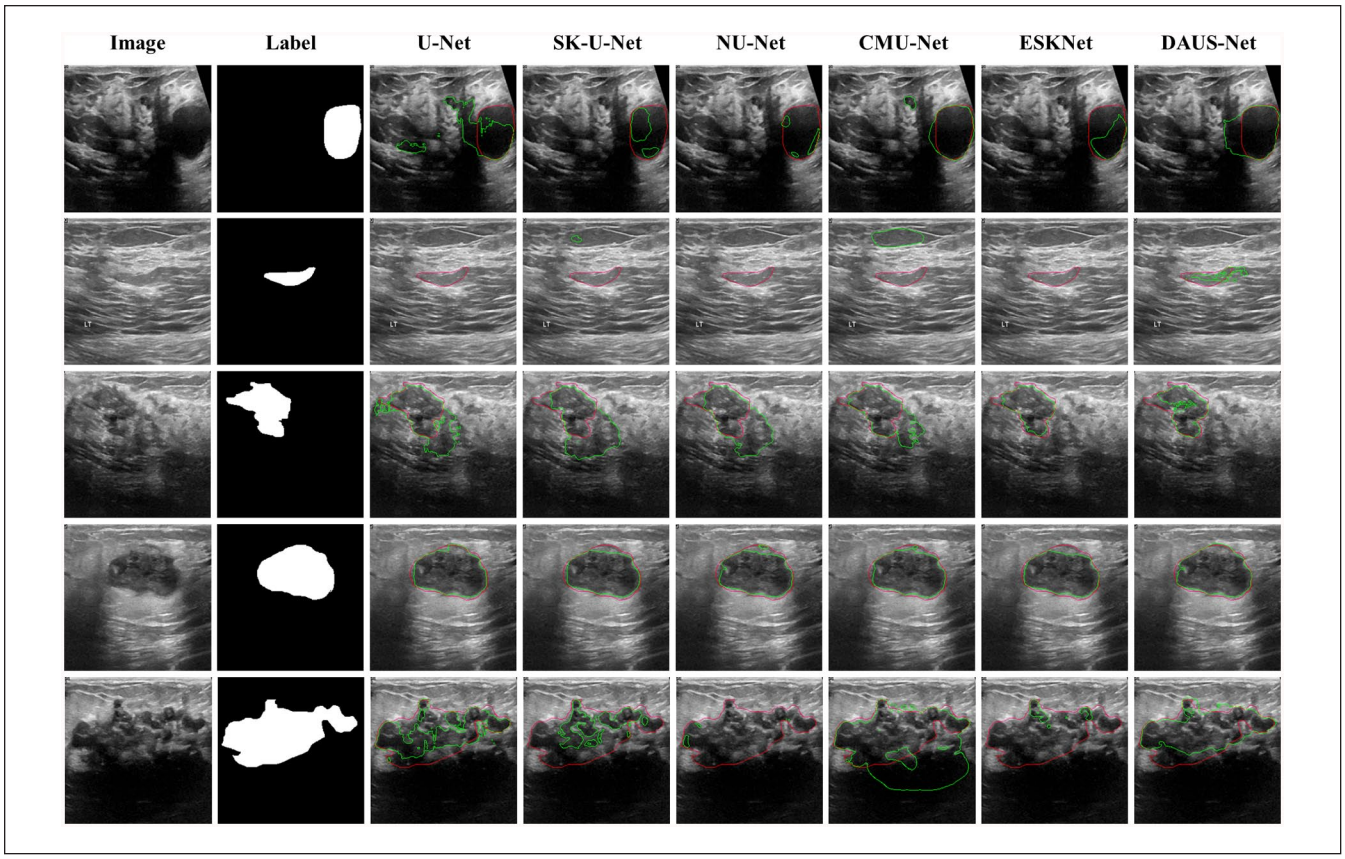

Figure 7. Predicted masks of U-Net, SK-U-Net, NU-Net, CMU-Net, ESKNet, and DAUS-Net trained/validated on Dataset B, BUS-BRA, BUS-UCLM and BLUI, and tested on BUSI. The red boundaries represent the label, and the green boundaries represent the predicted mask.

Table 2. Training/Validation on Dataset B and BUSI and Testing on BUS-BRA.

Model	Dice	IoU	Acc	Sens	Spec	AUC
U-Net	0.61 ± 0.05†↑	0.51 ± 0.04†↑	0.95 ± 0.00†	0.63 ± 0.07†↑	0.98 ± 0.00†	0.96 ± 0.01†
SK-U-Net	0.67 ± 0.05†↑	0.57 ± 0.05†↑	**0.96 ± 0.00**†	0.68 ± 0.06†↑	**0.99 ± 0.00**†	0.96 ± 0.01†
NU-Net	0.69 ± 0.08†	0.59 ± 0.08†	**0.96 ± 0.01**†	0.72 ± 0.12†↑	0.98 ± 0.00†	0.97 ± 0.01†
CMU-Net	0.65 ± 0.04†↑	0.55 ± 0.04†↑	**0.96 ± 0.00**†	0.64 ± 0.04†↑	**0.99 ± 0.00**†	0.97 ± 0.01†
ESKNet	0.72 ± 0.02†	0.62 ± 0.03†	**0.96 ± 0.00**	0.77 ± 0.04†	0.98 ± 0.01†	**0.98 ± 0.00**†
DAUS-Net	**0.76 ± 0.03**	**0.66 ± 0.03**	**0.96 ± 0.00**	**0.82 ± 0.05**	0.97 ± 0.01	**0.98 ± 0.01**

Bold values indicate the best performance for each metric.
† Denotes $q < 0.05$ and ↑ denotes Cohen's $d_x \geq 0.5$.

Table 3. Training/Validation on Dataset B and BUSI and Testing on BUS-UCLM.

Model	Dice	IoU	Acc	Sens	Spec	AUC
U-net	0.56 ± 0.04†↑	0.46 ± 0.03†↑	0.95 ± 0.00	0.61 ± 0.07†↑	0.98 ± 0.00†	0.95 ± 0.01†
SK-U-net	0.61 ± 0.05†	0.51 ± 0.04†	**0.96 ± 0.00**†	0.61 ± 0.07†↑	**0.99 ± 0.00**†	0.96 ± 0.01
NU-Net	0.65 ± 0.03	0.55 ± 0.02	**0.96 ± 0.00**†	0.67 ± 0.04†	**0.99 ± 0.00**†	0.97 ± 0.00†
CMU-Net	0.57 ± 0.04†↑	0.47 ± 0.04†↑	**0.96 ± 0.00**	0.56 ± 0.04†↑	**0.99 ± 0.00**†	0.96 ± 0.01†
ESKNet	0.64 ± 0.02†	0.54 ± 0.02†	**0.96 ± 0.00**	0.67 ± 0.07†	0.98 ± 0.01†	0.97 ± 0.01†
DAUS-Net	**0.67 ± 0.03**	**0.56 ± 0.03**	0.95 ± 0.01	**0.74 ± 0.04**	0.98 ± 0.01	0.97 ± 0.01

Bold values indicate the best performance for each metric.
† Denotes $q < 0.05$ and ↑ denotes Cohen's $d_x \geq 0.5$.

Table 4. Training/Validation on Dataset B and BUSI and Testing on BLUI.

Model	Dice	IoU	Acc	Sens	Spec	AUC
U-Net	0.75 ± 0.02†	0.64 ± 0.02†	0.93 ± 0.00†	0.75 ± 0.03†	0.98 ± 0.00	0.97 ± 0.00†
SK-U-Net	0.77 ± 0.02†	0.67 ± 0.03†	**0.94 ± 0.01**†	0.75 ± 0.02†	**0.99 ± 0.00**†	0.97 ± 0.01
NU-Net	0.78 ± 0.04†	0.69 ± 0.04	**0.94 ± 0.01**	0.77 ± 0.06	0.98 ± 0.00†	**0.98 ± 0.00**†
CMU-Net	0.73 ± 0.01†↑	0.63 ± 0.00†↑	**0.93 ± 0.00**†	0.70 ± 0.02†↑	0.99 ± 0.00†	0.97 ± 0.00†
ESKNet	0.77 ± 0.03†	0.68 ± 0.03†	**0.94 ± 0.01**†	0.75 ± 0.04†	0.98 ± 0.00†	0.97 ± 0.00†
DAUS-Net	**0.80 ± 0.02**	**0.70 ± 0.02**	**0.94 ± 0.00**	**0.79 ± 0.03**	0.98 ± 0.00	0.97 ± 0.00

Bold values indicate the best performance for each metric.
† Denotes $q < 0.05$ and ↑ denotes Cohen's $d_x \geq 0.5$.

Table 5. Training/Validation on Dataset B, BUS-BRA, BUS-UCLM, and BLUI and Testing on BUSI.

Model	Dice	IoU	Accuracy	Sens	Spec	AUC
U-Net	0.67 ± 0.01†	0.56 ± 0.00†	**0.95 ± 0.00**†	0.67 ± 0.03†	0.98 ± 0.00	0.95 ± 0.00†
SK-U-Net	0.69 ± 0.03†	0.59 ± 0.03†	**0.95 ± 0.00**†	0.68 ± 0.04†	0.98 ± 0.00†	0.95 ± 0.00†
NU-Net	0.71 ± 0.01	0.62 ± 0.01	**0.95 ± 0.00**†	0.70 ± 0.02†	0.98 ± 0.00†	0.95 ± 0.01†
CMU-Net	0.70 ± 0.00	0.61 ± 0.00	**0.95 ± 0.00**	0.70 ± 0.01	0.98 ± 0.00	0.95 ± 0.00
ESKNet	0.70 ± 0.01†	0.61 ± 0.01†	**0.95 ± 0.00**	0.67 ± 0.03†	**0.99 ± 0.00**†	**0.96 ± 0.00**†
DAUS-Net	**0.72 ± 0.01**	**0.63 ± 0.01**	**0.95 ± 0.00**	**0.72 ± 0.02**	0.98 ± 0.00	0.95 ± 0.00

Bold values indicate the best performance for each metric.
† Denotes $q < 0.05$ and ↑ denotes Cohen's $d_x \geq 0.5$.

Table 6. Computational Complexity and Efficiency of the Models.

Model	Params (M)	FLOPs-MAC (G)	FLOPs-2-op (G)	Latency mean (ms)	PeakMem (GB)	Throughput (FPS)
U-Net	31.0	48.3	109.3	5.2	0.2	190.8
SK-U-Net	4.9	4.8	10.3	6.8	0.2	146.3
NU-Net	77.0	40.1	80.2	14.2	0.7	70.5
CMU-Net	49.93	91.3	182.4	10.3	0.8	97
ESKNet	44.5	61.4	122.7	16.9	1	59.1
DAUS-Net	10.6	13.5	30.2	7.5	0.9	133.3

BUS-UCLM, and BLUI) and testing on BUSI. Although one might expect higher Dice and IoU scores with a larger number of training datasets, Table 5 shows that training on four datasets yielded lower performance than training on only two.

We benchmarked all models with an input size of $1 \times 1 \times 256 \times 256$, batch size 1, on an RTX A5000 GPU using fp32 precision, averaged over 100 runs. Table 6 presents the computational complexity and efficiency of the proposed and comparative models. We report FLOPs under two counting policies. In the FLOPs–MAC policy, one MAC is counted as a single operation, whereas in the FLOPs–2-op policy, a multiplication and an addition are counted separately (two operations per MAC).Ranked by FLOPs–MAC, the order is SK-U-Net < DAUS-Net < NU-Net < U-Net < ESKNet < CMU-Net. The FLOPs-2-op follows the same ranking as FLOPs–MAC. DAUS-Net achieves 13.55 g operations, 7.50 ms latency, 0.897 GB peak memory, and 133.3 FPS, offering the most balanced trade-off among computation, latency, and memory usage. While SK-U-Net requires fewer operations, its latency gains are limited. U-Net achieves the lowest latency but at higher computational cost, whereas ESKNet and CMU-Net exhibit greater latency and memory consumption. Although DAUS-Net records the second-highest peak memory, the absolute usage remains below 1 GB, well within typical clinical GPU budgets.

Discussion

Figure 8 presents the T-SNE visualization of feature maps extracted after the first, third, and fifth convolution blocks in the encoder. Blue squares correspond to DAUS-Net, and green circles correspond to U-Net. The visualized samples were obtained by inputting BUS-BRA data and selecting a specific channel from the output, with both models having been trained and validated on Dataset B and BUSI. As the feature maps progress through deeper layers, DAUS-Net, which uses instance normalization, gradually removes style-related variations, forming more compact clusters. In

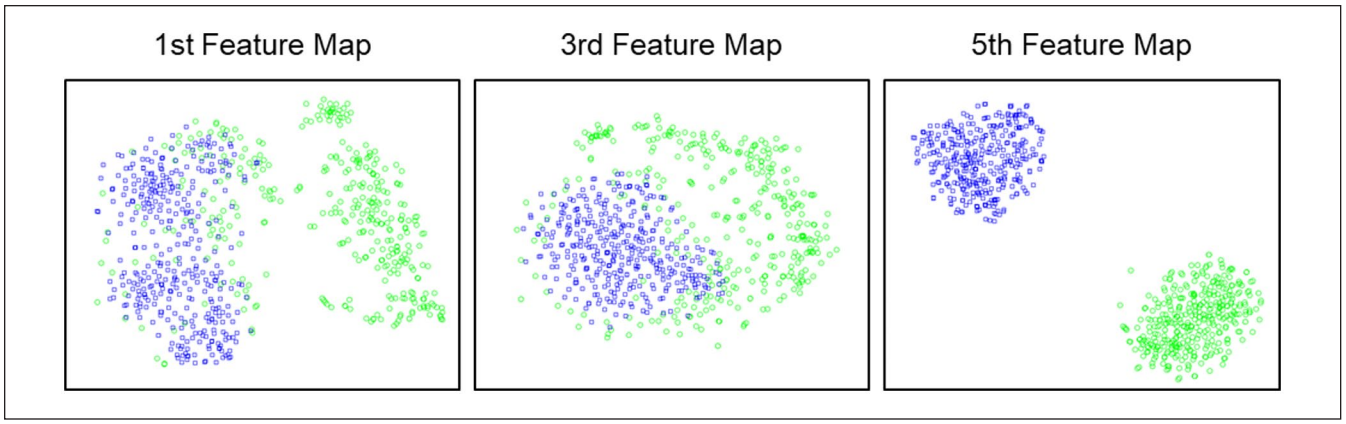

Figure 8. T-SNE visualization of feature maps obtained after the first, third, and fifth convolution blocks in the encoder. Blue squares represent feature map samples from DAUS-Net, while green circles represent feature map samples from U-Net. Each sample is derived by inputting BUS-BRA data and sampling a specific channel from the output.

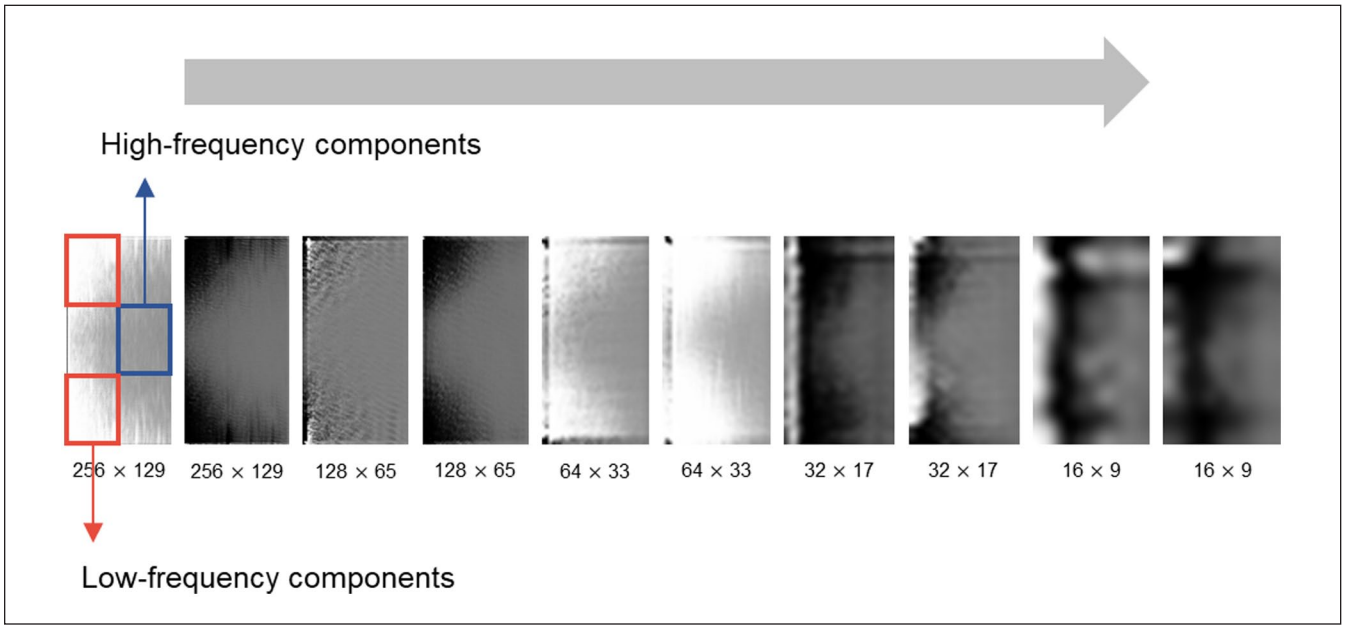

Figure 9. Frequency-space attention masks learned by DFF at each encoder stage of DAUS-Net. Masks are computed in the Fourier coordinate system and applied before the inverse FFT, thus functioning as content-and stage-adaptive frequency filters that selectively reweight spectral bands.

contrast, U-Net's feature maps remain scattered. By the fifth feature map, both DAUS-Net and U-Net coverage into a single distribution, likely due to the reduced feature map size of 16 × 16 pixels, which limits the expressive capacity of a single channel.

Figure 9 displays the spatial attention masks within the DFF module of the proposed model. In the mask on the far left, the red rectangular block represents LFC, while the blue rectangular block represents HFC. Brighter regions in the masks (values closer to 1) indicate areas where frequencies are allowed to pass through, whereas darker regions (values closer to 0) indicate areas of stronger filtering. From the far-left mask, it is evident that LFC are primarily passed through, while HFC are filtered out. As indicated by the gray arrows, the model alternates between filtering and passing frequencies as the layers progress, suggesting that it dynamically adjusts frequency components at each layer to effectively detect breast tumors.

Figure 10 compares the provided label with a label derived through intensity-based thresholding. In this figure, (a) displays the original image, (b) represents the provided label corresponding to the red boundaries and (c) shows a label generated by applying an intensity-based threshold corresponding to the blue boundaries. (e) illustrates the predicted mask from DAUS-Net, with Dice scores and IoU values compared against (d) for (b) and (f) for (c).

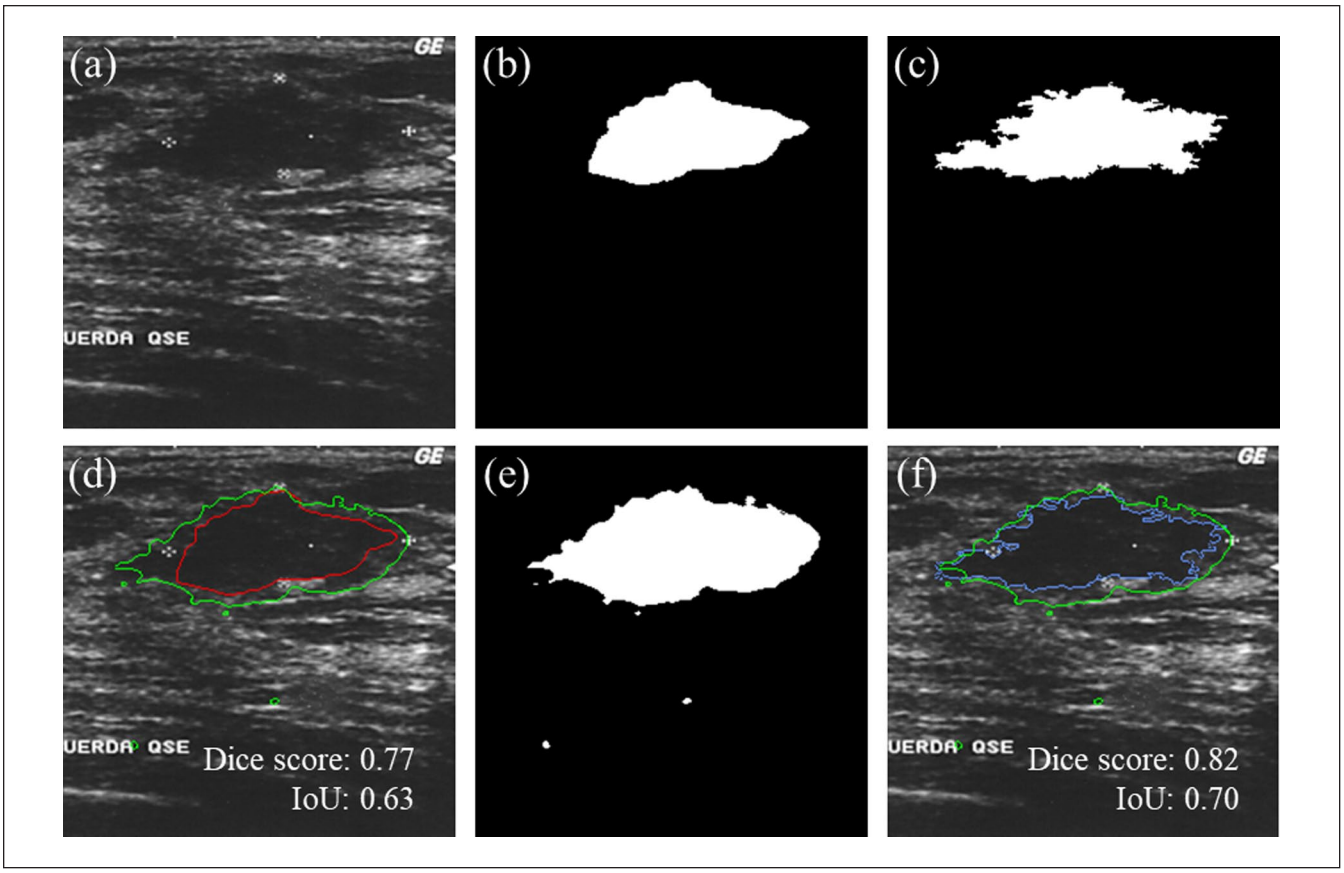

Figure 10. Comparison of the provided label and an intensity-based thresholding label. (a) Original image, (b) provided label, (c) threshold-generated label, (d) comparison of (b and e) predicted mask from DAUS-Net, (f) comparison of (c and e).

As noted in the introduction, tumor boundaries in breast ultrasound images are often ambiguous, resulting to high inter- and intra-rater variability during labeling.[23-26] This ambiguity makes it challenging to intuitively determine whether the tumor extends leftward, as in (c), or appears segmented, as in (b). Such uncertainty in tumor boundaries may partially account for the slightly lower Dice scores and IoU values. Nevertheless, as shown in (f), DAUS-Net demonstrates an intuitive ability to detect and segment breast tumors in a manner resembling human visual perception of tumor shapes. At the same time, we recognize that these interpretations may vary and emphasize the need for careful consideration, given the inherent ambiguity in labeling breast ultrasound images.

In typical segmentation tasks, the Dice score and IoU are widely regarded as key metrics for assessing how accurately a model delineates the target. In contrast, in the context of DG, achieving perfect segmentation is particularly challenging due to domain shifts and the inherent variability in medical imaging datasets. For tumor segmentation, simply detecting the tumor can be as important as precisely outlining its boundaries. In such cases, sensitivity becomes a critical metric, as it reflects the model's overall ability to identify tumors.

In this study, True Positive (TP) refers to cases where the model correctly identifies the tumor, while False Negative (FN) refers to cases where the model fails to identify the tumor, such as misclassifying the tumor as background. When models exhibit similar specificity, higher sensitivity indicates better tumor detection. From this perspective, results on the unseen test datasets show that the proposed model achieves notably higher sensitivity compared to the comparison models, indicating improved tumor detection. It also attains relatively higher Dice scores and IoU values, demonstrating robustness in detecting and segmenting tumors under the DG setting.

Clinical deployment requires stable ≥ 30 FPS, low per-frame latency, sufficient memory headroom alongside beamforming and UI, and predictable scaling across resolutions and devices. DAUS-Net meets these runtime constraints with 7.50 ms latency, 133.3 FPS, and 13.55 G operations, and its 0.897 GB peak memory fits within typical clinical GPU budgets. Even so, its segmentation accuracy has not fully generalized across all datasets, and the inherently ambiguous tumor margins make it premature to replace clinical judgment. We therefore position DAUS-Net as a radiologist-assist tool that, particularly for lesions with indistinct boundaries, provides a boundary confidence map and

exemplar-based suggestions to transparently indicate how the model would judge based on prior data.[52]

Conclusion

In this study, we proposed DAUS-Net to address the challenges of DG in breast ultrasound segmentation. The model incorporates a DFF module in the encoder and replaces BN with IN to enhance robustness across domains. The DFF module and IN work synergistically to mitigate domain shifts by preserving the discriminative information needed to distinguish breast tumors from the background while removing stylish variations. Experiments conducted on five public datasets—Dataset B, BUSI, BUS-BRA, BUS-UCLM, and BLUI—showed that the proposed model achieved superior Dice, IoU, and sensitivity relative to the baseline models. By leveraging frequency-specific features of breast tumors, DAUS-Net reliably detected tumor presence and segmented their boundaries, even on unseen datasets. Given the inherent ambiguity in breast ultrasound images, the model exhibits a tendency to adopt human-like decision-making processes when addressing unclear tumor boundaries. Although the proposed model demonstrates superior performance compared to the baselines, the results vary depending on how the training/validation and test datasets are configured. Therefore, rather than positioning DAUS-Net as a stand-alone solution, we regard it as a radiologist-assist tool that provides exemplar-based suggestions and boundary confidence cues to support clinical decision-making in cases of ambiguous tumor margins.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was supported by the National Research Foundation of Korea (NRF) funded by the Korea government (MSIT) (NRF-RS-2023-00211941), the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (RS-2025-25420118, RS-2025-02305555), and the DGIST R&D Program of the Ministry of Science and Information and Communications Technology (ICT) (25-DRJoint-02).

ORCID iD

Jaesok Yu https://orcid.org/0000-0001-8157-1718

References

1. Noble JA, Boukerroui D. Ultrasound image segmentation: A survey. IEEE Trans Med Imaging. 2006;25(8):987-1010.
2. Jardim SM, Figueiredo MA. Segmentation of fetal ultrasound images. Ultrasound Med Biol. 2005;31(2):243-50.
3. Xian M, Zhang Y, Cheng HD, Xu F, Zhang B, Ding J. Automatic breast ultrasound image segmentation: A survey. Pattern Recognit. 2018;79:340-55.
4. Huang Q, Luo Y, Zhang Q. Breast ultrasound image segmentation: a survey. Int J Comput Assist Radiol Surg. 2017;12(3):493-507.
5. Chang RF, Wu WJ, Moon WK, Chen DR. Automatic ultrasound segmentation and morphology based diagnosis of solid breast tumors. Breast Cancer Res Treat. 2005;89(2):179-85.
6. Bosch JG, Mitchell SC, Lelieveldt BP, Nijland F, Kamp O, Sonka M, et al. Automatic segmentation of echocardiographic sequences by active appearance motion models. IEEE Trans Med Imaging. 2002;21(11):1374-83.
7. Chang JM, Moon WK, Cho N, Kim SJ. Breast mass evaluation: Factors influencing the quality of US elastography. Radiology. 2011;259(1):59-64.
8. Vakanski A, Xian M, Freer PE. Attention-enriched deep learning model for breast tumor segmentation in ultrasound images. Ultrasound Med Biol. 2020;46(10):2819-33.
9. Amiri M, Brooks R, Behboodi B, Rivaz H. Two-stage ultrasound image segmentation using U-net and test time augmentation. Int J Comput Assist Radiol Surg. 2020;15(6):981-8.
10. Byra M, Jarosik P, Szubert A, Galperin M, Ojeda-Fournier H, Olson L, et al. Breast mass segmentation in ultrasound with selective kernel U-net convolutional neural network. Biomed Signal Process Control. 2020;61:102027.
11. Tong Y, Liu Y, Zhao M, Meng L, Zhang J. Improved U-net MALF model for lesion segmentation in breast ultrasound images. Biomed Signal Process Control. 2021;68:102721.
12. Chen G, Dai Y, Zhang J. C-Net: cascaded convolutional neural network with global guidance and refinement residuals for breast ultrasound images segmentation. Comput Methods Programs Biomed. 2022;225:107086.
13. Ning Z, Zhong S, Feng Q, Chen W, Zhang Y. SMU-Net: saliency-guided morphology-aware U-net for breast lesion segmentation in ultrasound image. IEEE Trans Med Imaging. 2022;41(2):476-90.
14. Chen G, Li L, Dai Y, Zhang J, Yap MH. AAU-Net: an adaptive attention U-net for breast lesions segmentation in ultrasound images. IEEE Trans Med Imaging. 2023;42(5):1289-300.
15. Chen G, Li L, Zhang J, Dai Y. Rethinking the unpretentious U-net for medical ultrasound image segmentation. Pattern Recognit. 2023;142:109728.
16. Tang F, Wang L, Ning C, Xian M, Ding J. CMU-NeT: a strong Convmixer-based medical ultrasound image segmentation network. In: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI), Cartagena, Colombia, 2023, pp. 1-5.
17. Chen G, Zhou L, Zhang J, Yin X, Cui L, Dai Y. ESKNet: an enhanced adaptive selection kernel convolution for ultrasound breast tumors segmentation. Expert Syst Appl. 2024;246:123265.
18. Liu Z, Huang X, Yang X, Gao R, Li R, Zhang Y. Generalize Ultrasound Image Segmentation Via Instant and Plug & Play. Shenzhen: Medical UltraSound Image Computing (MUSIC) Lab, Shenzhen University, China Department of Ultrasound, Luohu People's Hospital, Shenzhen; 2021. pp.419-23. China School of Biological Sciences and M.
19. Kaur B, Lemaˆ P, Mehta R, Arbel T. Improving pathological structure segmentation via transfer learning across diseases. In: Wang Q, et al., eds. Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and

Imperfect Data. DART MIL3ID 2019. Lecture Notes in Computer Science, vol 11795. Cham: Springer, Cham, 2019.
20. Hamed Mozaffari M, Lee W-S. Domain adaptation for ultrasound tongue contour extraction using transfer learning: A deep learning approach. J Acoust Soc Am. 2019;146:EL431-7.
21. Karimi D, Dou H, Warfield SK, Gholipour A. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. Med Image Anal. 2020;65:101759.
22. Zhou K, Liu Z, Qiao Y, Xiang T, Loy CC. Domain generalization: A survey. IEEE Trans Pattern Anal Mach Intell. 2023;45(4):4396-415.
23. Chisholm RA, Stenning S, Hawkins TD. The accuracy of volumetric measurement of high-grade gliomas. Clin Radiol. 1989;40(1):17-21.
24. Menze BH, Jakab A, Bauer S, Kalpathy-Cramer J, Farahani K, Kirby J, et al. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS. IEEE Trans Med Imaging. 2015;34(10):1993-2024.
25. Rawashdeh M, Lewis S, Zaitoun M, Brennan P. Breast lesion shape and margin evaluation: BI-RADS based metrics understate radiologists' actual levels of agreement. Comput Biol Med. 2018;96:294-8.
26. Huang K, Zhang Y, Cheng HD, Xing P, Zhang B. Semantic segmentation of breast ultrasound image with fuzzy deep learning network and breast anatomy constraints. Neurocomputing. 2021;450:319-35.
27. Gu R, Wang G, Lu J, Zhang J, Lei W, Chen Y, et al. CDDSA: Contrastive domain disentanglement and style augmentation for generalizable medical image segmentation. Med Image Anal. 2023;89:102904.
28. Kim H, Shin Y, Hwang D. DiMix: Disentangle-and-Mix Based Domain Generalizable Medical Image Segmentation, In Greenspan H, et al., eds. Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. Lecture Notes in Computer Science, vol 14222. Cham: Springer.
29. Jin X, Lan C, Zeng W, Chen Z. Style normalization and restitution for domain generalization and adaptation. IEEE Trans Multimedia. 2022;24:3636-51.
30. Pan X, Luo P, Shi J, Tang X. Two at once: enhancing learning and generalization capacities via IBN-Net. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 11208 LNCS, pp. 484-500, 2018, doi:10.1007/978-3-030-01225-0_29.
31. Chen Y, Zhang H, Wang Y, Peng W, Zhang W, Wu QMJ, et al. D-BIN: A generalized disentangling batch instance normalization for domain adaptation. IEEE Trans Cybern. 2023;53(4):2151-63.
32. Choi S, Kim T, Jeong M, Park H, Kim C. Meta batch-instance normalization for generalizable person re-identification. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Los Alamitos, CA, USA. IEEE Computer Society, 2021, pp. 3425-35.
33. Lin S, Zhang Z, Huang Z, Lu Y, Lan C, Chu P, et al. Deep frequency filtering for domain generalization. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Los Alamitos, CA, USA, 2023, pp. 11797-807. IEEE Computer Society.
34. Wang H, Wu X, Huang Z, Xing EP. High-frequency component helps explain the generalization of convolutional neural networks. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Los Alamitos, CA, USA. IEEE Computer Society, 2020, pp. 8684-94.
35. Xu ZJ. Understanding training and generalization in deep learning by Fourier analysis. no. 2002, 2018 [Online]. http://arxiv.org/abs/1808.04295
36. Navab N, Hornegger J, Wells WM, Frangi AF. U-Net: convolutional networks for biomedical image segmentation. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 9351, no. Cvd, pp.12-20, 2015, doi:10.1007/978-3-319-24574-4.
37. Woo S, Park J, Lee J, Kweon IS. CBAM: convolutional block attention module. In: Computer Vision – ECCV 2018, Proceedings of the 15th European Conference on Computer Vision, Munich, Germany, September 8–14, 2018, Lecture Notes in Computer Science, vol. 11211. Cham: Springer, 2018. pp. 3-19.
38. Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. 32nd Int. Conf. Mach. Learn. ICML 2015. 2015;1:448-56.
39. Ulyanov D, Vedaldi A, Lempitsky V. Instance normalization: The missing ingredient for fast stylization. no. 2016, 2016, [Online]. http://arxiv.org/abs/1607.08022
40. de Boer PT, Kroese DP, Mannor S, Rubinstein RY. A tutorial on the cross-entropy method. Ann Oper Res. 2005;134(1):19-67.
41. Yap MH, Pons G, Martí J, Ganau S, Sentís M, Zwiggelaar R, et al. Automated breast ultrasound lesions detection using convolutional neural networks. IEEE J Biomed Health Inform. 2018;22(4):1218-26.
42. Al-Dhabyani W, Gomaa M, Khaled H, Fahmy A. Dataset of breast ultrasound images. Data Brief. 2020;28:104863.
43. Gómez-Flores W, Gregorio-Calas MJ, Coelho de Albuquerque Pereira W. BUS-BRA: A breast ultrasound dataset for assessing computer-aided diagnosis systems. Med Phys. 2024;51(4):3110-23.
44. Vallez N, Bueno G, Deniz O, Rienda MA, Pastor C. BUS-UCLM: Breast ultrasound lesion segmentation dataset. Sci Data. 2025;12(1):242-8.
45. Abbasian Ardakani A, Mohammadi A, Mirza-Aghazadeh-Attari M, Acharya UR. An open-access breast lesion ultrasound image database: applicable in artificial intelligence studies. Comput Biol Med. 2023;152:106438.
46. García-Alonso CR, Pérez-Naranjo LM, Fernández-Caballero JC. Visualizing data using t-SNE. Ann Oper Res. 2014;219(1):187-202.
47. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: an imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems 32 (NeurIPS 2019), Red Hook, NY, USA. Curran Associates Inc., 2019.
48. Müller D, Soto-Rey I, Kramer F. Towards a guideline for evaluation metrics in medical image segmentation. BMC Res Notes. 2022;15(1):210-8.
49. Woolson RF. Wilcoxon Signed-Rank Test. In: D'Agostino RB, Sullivan L, Massaro J, eds. Wiley Encyclopedia of Clinical Trials. doi:10.1002/9780471462422.eoct979
50. February R. The control of the false discovery rate in multiple testing under dependency by Yoav Benjamini 1 and Daniel Yekutieli 2. Ann Stat. 2001;29(4):1165-88.
51. Lakens D. Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAS. Front Psychol. 2013;4:1-12.
52. Wu N, Phang J, Park J, Shen Y, Huang Z, Zorin M, et al. Deep neural networks improve radiologists' performance in breast cancer screening. IEEE Trans Med Imaging. 2020;39(4):1184-94.