Semi-Supervised Breast Ultrasound Segmentation via Text-Guided Foundation Model

Šimon Freivald*
Supervised by: Igor Jánoš†

Faculty of Informatics and Information Technologies
Slovak University of Technology
Bratislava / Slovakia

Abstract

Breast cancer remains the leading cause of cancer-related mortality among women worldwide. Ultrasound imaging has become a preferred modality for breast screening due to its non-invasive nature, absence of ionizing radiation, and low cost. However, accurate segmentation of breast lesions in ultrasound images remains challenging. Furthermore, the scarcity of pixel-level annotated data in the medical domain significantly limits the training of deep learning models. In this paper, we propose a transformer-based framework for breast tumor segmentation that combines the Universal Ultrasound Foundation Model (USFM) for visual feature extraction with a CLIP text encoder for semantic guidance via cross-attention. To address the data scarcity challenge, we design a teacher-student training scheme that effectively leverages both strongly annotated data with pixel-precise masks and a substantially larger set of weakly annotated data with only bounding box labels. We aggregate a comprehensive training set of strongly annotated and weakly annotated ultrasound images from multiple public datasets, enriched by data from multiple organs. The proposed method is evaluated on the widely used BUSI and BUS UC benchmark datasets.

Keywords: ultrasound segmentation, breast cancer, semi-supervised learning, foundation model, text guidance

1 Introduction

Breast cancer is the most common cause of cancer-related mortality worldwide, responsible for 670,000 deaths in 2022 [2]. Early detection through regular screening remains critical for reducing mortality. Among non-invasive imaging modalities, ultrasound has emerged as a highly promising technique due to its accessibility, affordability, real-time imaging capability, and absence of ionizing radiation [7, 35]. Ultrasound is particularly effective at differentiating cysts from solid lesions even in dense breast tissue [21], making it a first-line choice for breast cancer diagnosis.

Computer-aided diagnosis (CAD) systems have been developed to assist radiologists in localizing and classifying breast lesions, with image segmentation serving as a foundational step in the CAD pipeline. Accurate tumor segmentation provides the basis for subsequent feature extraction and classification of benign versus malignant lesions [25, 35]. In clinical practice, manual segmentation is tedious, time-consuming, and subject to inter-observer variability, which motivates the development of reliable automated methods [14].

Convolutional neural networks (CNNs) have been widely adopted for medical image analysis. However, their inherently local receptive fields limit the capture of global contextual information, which is critical when lesions occupy only a small portion of the image [7, 12]. Recently, transformer architectures have demonstrated the ability to effectively capture long-range dependencies through self-attention mechanisms, achieving state-of-the-art results in medical image segmentation [11]. Nevertheless, most transformer-based methods require pre-training on large-scale datasets, which poses a significant barrier in the medical domain where annotated data is scarce [25].

In this work, we address these challenges with two key contributions. First, we propose an architecture that combines a domain-specific ultrasound foundation model (USFM) [16] with a CLIP text encoder [23] through cross-attention, enabling text-guided visual feature enrichment for segmentation. Second, we design a teacher-student training framework that jointly leverages strongly annotated data (pixel-precise masks) and a substantially larger pool of weakly annotated data (bounding box labels), thereby alleviating the dependence on expensive pixel-level annotations.

2 Related Work

CNN-based breast ultrasound segmentation. Early deep learning approaches to medical image segmentation were dominated by convolutional neural networks, with U-Net [24] establishing the foundational encoder-decoder architecture that remains influential to this day. Subsequent

*freivaldsimon@gmail.com
†igor.janos@innovatrics.com

work tailored this paradigm to breast ultrasound by introducing multi-scale feature aggregation, boundary-aware supervision, and attention-gated skip connections [26]. Despite strong empirical results, CNN-based architectures suffer from an inherently limited receptive field, restricting their ability to reason about global context, a limitation particularly pronounced in ultrasound, where acoustic shadowing and speckle noise cause lesion boundaries to be ambiguous across large spatial extents [7, 12].

Transformer-based approaches. Vision transformers capture long-range dependencies inaccessible to local convolutions, which is especially valuable when tumors exhibit heterogeneous echo patterns relative to surrounding tissue [11]. For breast ultrasound, CSwin-Pnet [33] introduced cross-shaped window attention for multi-scale spatial context, while HA-Net [4] proposed a hierarchical attention mechanism targeting lesion boundaries. These methods consistently outperform convolutional baselines but are typically trained from scratch on small in-domain datasets, limiting generalization to unseen acquisition conditions.

Foundation models and prompt-based segmentation. Large-scale pre-trained foundation models have opened a new paradigm for medical image analysis. SAM [17] demonstrated impressive zero-shot segmentation via sparse geometric prompts, with medical adaptations such as MedSAM [20] and BUSSAM [29] closing the domain gap for clinical and ultrasound images respectively. However, SAM-derived models lack modality-specific inductive biases for ultrasound. USFM [16] addresses this directly as a ViT-B foundation model pre-trained on approximately 2.2 million multi-organ ultrasound images via self-supervised learning. Concurrently, vision-language models such as CLIP [23] have enabled few-shot recognition and weakly supervised segmentation in medical imaging [19]. Our work combines the ultrasound-specific representations of USFM with CLIP-based semantic guidance through cross-attention, extending supervision to bounding-box annotations via a teacher-student scheme [27].

3 Method

We propose a transformer-based segmentation framework that combines a frozen ultrasound-specific visual backbone with semantic text guidance and a teacher–student training scheme for mixed supervision. Given an input image $\mathbf{x} \in \mathbb{R}^{3\times256\times256}$ and an organ-class text prompt p, the model produces a binary segmentation map $\hat{\mathbf{y}} \in [0,1]^{256\times256}$. The pipeline consists of four components: (i) a frozen USFM ViT backbone for multi-scale visual feature extraction, (ii) a frozen CLIP text encoder with a lightweight adapter for semantic embedding, (iii) per-scale cross-attention modules that fuse text and visual tokens, and (iv) a hierarchical multi-scale decoder that produces the final segmentation mask. During training, a teacher network is maintained as an exponential moving average of the student, which generates pseudo-labels for weakly annotated samples.

3.1 Preprocessing and Input Representation

Most prior work processes single-channel grayscale images by replicating the channel three times to meet the input requirements of pretrained feature extractors, which are typically designed for RGB images. Rather than duplicating identical channels, we leverage this three-channel structure to incorporate complementary intensity representations. Specifically, each raw ultrasound image is transformed into a three-channel tensor composed of: (i) the original grayscale image, (ii) a contrast-stretched version, and (iii) a contrast-limited adaptive histogram equalization (CLAHE)-enhanced version. This strategy enables improved representation of intensity variations and mitigates the wide dynamic-range variability characteristic of ultrasound imaging, while preserving compatibility with standard pretrained architectures. Notably, this approach does not require any architectural modifications. All input images are resized to 256×256 pixels prior to training.

3.2 USFM Backbone

The visual backbone is the HViT encoder from USFM [16]. It is implemented as a 12-layer Vision Transformer (ViT-B) with an embedding dimension of 768 and a patch size of 16, resulting in a 16×16 grid of patch tokens for a 256×256 input image. The USFM model was pretrained by self-supervised learning on the 3M-US dataset, which comprises 2,187,915 unlabeled ultrasound images of twelve common organs used in routine human body screening. The dataset aggregates data from multiple medical centers and diagnostic devices worldwide, substantially enhancing the robustness and generalization capability of the pretrained model. In this work, the pretrained USFM encoder is employed for initial feature extraction from input ultrasound images. Intermediate representations are extracted from layers 10, 11, and 12. These activations are projected to a common channel dimension and reshaped into three multi-scale feature maps with spatial resolutions of 64×64, 32×32, and 16×16, denoted as F_1, F_2, and F_3, respectively. These feature maps serve as the input representations for subsequent processing stages. During training, all backbone parameters remain frozen except the first layer, which is unfrozen to allow the backbone to adapt to the enriched three-channel input representation.

3.3 CLIP Text Encoder and Adaptation

Organ-specific prompts are embedded using a frozen `openai/clip-vit-base-patch16` CLIPTextModel [23]. The primary objective of this component is to obtain a numerical representation of textual descriptions corresponding to the input ultrasound image. This description can be interpreted as a high-level characterization of the image, including the imaged organ and the potential presence of pathological findings.

Since the CLIP text encoder was originally pretrained on large-scale, predominantly non-medical corpora, a domain gap exists between its learned language representations and the specialized semantics of ultrasound imaging. To mitigate this mismatch, we introduce a lightweight **text adapter** placed after the frozen CLIP encoder. The adapter consists of a two-layer multi-layer perceptron (MLP) with GELU activation and LayerNorm, mapping the 512-dimensional CLIP token embeddings to the 256-dimensional channel space of the visual feature maps. This adapter is the only trainable component on the text branch and is optimized end-to-end, enabling better alignment of textual embeddings with the medical imaging domain.

Each organ category (breast, thyroid, liver, and ovary) is associated with ten descriptive prompts capturing sonographic appearance, echo pattern, and anatomical context. For example:

- *"Breast ultrasound showing tissue distortion and irregular internal echoes."*
- *"Sonographic view of thyroid lobe with mixed echogenicity and internal reflections."*

This design is particularly important as our method leverages auxiliary datasets originating from organs other than the breast, which constitutes the primary object of investigation. During both training and inference, a prompt is sampled uniformly at random from the corresponding organ-specific pool. This strategy acts as a language-level augmentation mechanism and reduces overfitting to a single prompt formulation.

3.4 Per-Scale Cross-Attention Fusion

At each feature scale $i \in \{1,2,3\}$, we apply a **TextCrossAttention** module to fuse visual features with the corresponding text representation. The goal of this module is to enrich visual feature maps with semantic information derived from the textual description of the input ultrasound image.

Let $F_i \in \mathbb{R}^{B\times C_i\times H_i\times W_i}$ denote the visual feature map at scale i, and let $T \in \mathbb{R}^{B\times T\times d_t}$ represent the adapted text token embeddings. First, F_i is projected to a shared latent dimension d using a 1×1 convolution. The spatial dimensions are then flattened into a sequence of visual tokens $X_i \in \mathbb{R}^{B\times N\times d}$, where $N = H_i W_i$.

In parallel, the text tokens are linearly projected to the same latent space. Cross-modal interaction is performed using multi-head attention, where visual tokens serve as *queries* and text tokens act as *keys* and *values*:

$$X_i' = \mathrm{MHA}(Q = X_i,\, K = \tilde{T},\, V = \tilde{T}). \tag{1}$$

The attention output is added to the original visual tokens via a residual connection, followed by layer normalization. A position-wise feed-forward network with GELU activation is then applied, again with residual connection and normalization. The refined tokens are reshaped back to the original spatial layout and projected to the initial channel dimension, yielding the text-conditioned feature map F_i'.

This design keeps the visual backbone frozen while enabling each feature scale to be selectively modulated by high-level semantic information from the language prompt. The resulting enriched feature maps are subsequently used for more accurate tumor segmentation in ultrasound images.

3.5 Multi-Scale Decoder

The three text-conditioned feature maps F_1', F_2', F_3' are processed by a **ThreeScaleDecoder** inspired by UPerNet [31] and feature pyramid networks. All three maps are first projected to a unified channel dimension via 1×1 convolutions and refined by two consecutive 3×3 convolutional layers with GroupNorm and ReLU. Fusion proceeds hierarchically from deep to shallow: the deepest map (16×16) is bilinearly upsampled and gated-fused with the intermediate map (32×32) via a learned sigmoid weighting, then the result is fused with the shallowest map (64×64) in the same manner. The final feature map is projected to a single-channel logit map and upsampled to 256×256.

3.6 Teacher–Student Training

We adopt a teacher–student scheme following Mean Teacher [27]. The *student* network is trained with gradient descent; the *teacher* is an exponential moving average (EMA) of the student with a dynamically ramped decay coefficient:

$$\alpha(t) = \alpha_{\text{start}} + (\alpha_{\text{end}} - \alpha_{\text{start}}) \cdot \frac{t}{T-1}, \tag{2}$$

where t is the current epoch, T is the total epoch count, $\alpha_{\text{start}} = 0.99$, and $\alpha_{\text{end}} = 0.999$. Early in training, the teacher tracks the student quickly; as the student stabilises, the teacher accumulates a longer history, producing smoother pseudo-labels.

3.6.1 Strong Branch.

Pixel-precise masks are available for fully annotated samples (flag $= 0$). The student predicts logits $\hat{\ell}$ and the strong loss is:

$$\mathscr{L}_{\text{strong}} = \mathscr{L}_{\text{DiceFocal}}(\hat{\ell}, y), \tag{3}$$

using MONAI's DiceFocalLoss [6] with equal Dice and focal weights, $\gamma = 2$.

3.6.2 Weak Branch

For weakly annotated samples (flag $= 1$), only bounding boxes are provided, without pixel-wise segmentation. To leverage such weak annotations, we adopted the teacher-student framework. The teacher processes the clean input

image and outputs soft pseudo-label probabilities $\tilde{p}$, providing an initial estimate of likely foreground and background regions.

Hard pseudo-labels are then constructed using confidence thresholds:

$$\tilde{y}_{\text{fg}} = [\tilde{p} > 0.75], \tag{4}$$

$$\tilde{y}_{\text{bg}} = \min([\tilde{p} < 0.25] + (1 - y_{\text{box}}), 1), \tag{5}$$

where y_{box} denotes the bounding-box mask. Pixels inside the bounding box that are neither confidently foreground nor confidently background are treated as an *ignore region* ($\omega = 0$). These pixels are excluded from gradient computation, preventing the student from learning from ambiguous or noisy labels.

The student network then processes a perturbed version of the same image using additional augmentation. Supervision is applied through two complementary losses:

Weighted Binary Cross-Entropy (wBCE) Confident pseudo-labels are used to compute a weighted BCE loss:

$$\mathscr{L}_{\text{wBCE}} = \frac{1}{|\Omega|} \sum_{(i,j)\in\Omega} \text{BCE}(\hat{\ell}_{ij}, \tilde{y}_{\text{fg},ij}), \tag{6}$$

where Ω denotes the set of pixels outside the ignore region. This ensures the student focuses on reliable foreground and background regions while ignoring uncertain areas.

Box Projection Loss Since weak annotations provide only bounding boxes, we enforce consistency between the predicted mask and the box along each axis. Specifically, the maximum projection of the predicted mask along the horizontal and vertical axes should match the corresponding bounding-box projections [28]:

$$\begin{aligned}\mathscr{L}_{\text{proj}} = \text{BCE}\Big(\max_j \hat{\ell}_{i\cdot}, \max_j y_{\text{box},i\cdot}\Big) \\ + \text{BCE}\Big(\max_i \hat{\ell}_{\cdot j}, \max_i y_{\text{box},\cdot j}\Big).\end{aligned} \tag{7}$$

This auxiliary loss guides the student to produce masks consistent with the known bounding-box geometry, even in regions without confident pseudo-labels.

Overall Weak Supervision The total loss for weakly labeled samples combines the two terms:

$$\mathscr{L}_{\text{weak}} = \mathscr{L}_{\text{wBCE}} + \lambda \mathscr{L}_{\text{proj}}, \tag{8}$$

with $\lambda = 0.5$ in our experiments. This strategy allows the model to learn from weak annotations effectively by (i) leveraging confident pseudo-labels, (ii) ignoring uncertain regions to reduce noise, and (iii) enforcing spatial consistency through box projection constraints.

3.6.3 Total Loss and Dynamic Scheduling.

The total loss blends the two branches proportionally to the batch composition:

$$\mathscr{L} = \frac{n_s}{n_s + n_w} \mathscr{L}_{\text{strong}} + \frac{n_w}{n_s + n_w} \lambda(t) \mathscr{L}_{\text{weak}}, \tag{9}$$

where n_s, n_w are the per-batch strong and weak counts, and $\lambda(t)$ linearly ramps from 0.1 to 0.5 over training. The weak branch is not activated for the first $K = 5$ warmup epochs, preventing early corruption from poor teacher pseudo-labels.

4 Datasets and Data Integrity

In medical imaging, the scarcity of high-quality annotated datasets is a long-standing issue, particularly in breast ultrasound. To mitigate limited training data, we employ both fully and weakly annotated datasets, which serve complementary roles in model training.

Fully annotated datasets provide pixel-precise masks that are crucial for learning fine-grained tumor boundaries. In our work, we aggregate multiple fully annotated public datasets across different organs to enrich diversity:

4.1 Strongly Annotated Datasets

- **BUS-BRA** [8]: 1,875 breast ultrasound images from 1,064 patients aged 16–89 years. Manual pixel-wise annotations were performed for benign (722) and malignant (342) tumors.
- **BrEaST** [22]: 256 patients (154 benign, 98 malignant, 4 normal), with manual segmentation performed by five radiologists in Poland (2019–2022).
- **BUS-UCLM** [30]: 683 images from 38 patients (174 benign, 90 malignant, 419 normal), annotated independently by two radiologists.
- **BUSI-WHU** [13]: 927 breast ultrasound images (560 benign, 367 malignant), with annotations verified by multiple radiologists.
- **BUID** [3]: 232 images (123 malignant, 109 benign), annotated based on histopathological reports and expert review.

4.1.1 Supporting Strongly Annotated Datasets

To improve generalization beyond breast ultrasound, we include strongly annotated supporting datasets from other organs:

- **TN3K+TG3K** [10][9]: 6,463 thyroid ultrasound images of the gland and nodules.
- **LUS** [32]: 735 liver images annotated for malignant, benign, and normal tissue.

- **MMOTU** [34]: 1,639 ovarian ultrasound images from 294 patients.

4.2 Weakly Annotated Datasets

The largest portion of our dataset comes from weakly annotated images, which provide only bounding boxes. This includes:

- **CVA / VideoBUS** [18]: 188 breast ultrasound videos with 25,272 frames annotated with per-frame bounding boxes (113 malignant, 75 benign). These weak labels enable scalable training using our weak branch, as described in Section 3.6.2.

4.3 Test Datasets

For evaluation, we adopt fully annotated datasets that are widely used in the literature:

- **BUSI** [1]: 780 breast ultrasound images from 600 patients, manually segmented. Despite popularity, this dataset suffers from severe duplication issues, which can bias reported metrics (see Section 4.4).
- **BUS UC** [15]: 811 images from UltrasoundCase portal (358 benign, 453 malignant), used as a clean, independent benchmark to assess true generalization.

4.4 BUSI Data-Integrity Analysis

The BUSI dataset [1] has become the de facto baseline benchmark for breast ultrasound segmentation. Despite its popularity, prior work [5] identified a serious issue with duplicate images. Using an algorithm to detect duplicates, it was found that, out of the original 780 images - 5 images have quadruple copies, 22 images have triple copies and 122 images have double copies. Some of these duplicates differ slightly in their annotations. Such duplication can severely bias model evaluation: randomly splitting the dataset may place near-identical images in both training and validation sets. As a result, models can appear to perform exceptionally well on validation images that are effectively duplicates of training samples, leading to overestimated metrics and potentially overfitting.

Based on these findings, we adopt BUS UC [15] as our primary hold-out test set and report results on BUSI only for legacy comparison. Using a clean, independent benchmark ensures that evaluation metrics reflect true generalization and are not artificially inflated by duplicated samples.

5 Experiments

5.1 Training Details

All models are trained for 200 epochs with AdamW ($\beta = (0.9, 0.999)$, weight decay 10^{-4}) using a two-tier learning rate: $\mathrm{lr} = 10^{-5}$ for the backbone (first-layer) and $\mathrm{lr} = 10^{-4}$ for all downstream modules. A 5-epoch linear warmup is followed by cosine annealing to $\eta_{\min} = 10^{-6}$. Training uses batch size 64 (mixed strong and weak samples), bfloat16 AMP, and gradient clipping at norm 1.0. All experiments run on a single NVIDIA RTX 4090. Geometric augmentation includes random horizontal and vertical flips, rotation (± 15), and perspective transform. Weakly annotated samples are additionally perturbed with multiplicative speckle noise, Gaussian noise, random blur or sharpening, and gamma variation to simulate acquisition variability.

5.2 Comparison on the BUSI Dataset

We first compare our method against recent state-of-the-art approaches on the BUSI dataset. Table 1 reports the Dice and Jaccard scores alongside the training and validation split ratios used by each method.

Table 1: Comparison with state-of-the-art methods on the BUSI dataset. Train/Val denotes the proportion of data used for training and validation, respectively.

Method	Train	Val	DSC (%)	IoU (%)
CSwin-Pnet [33]	80%	20%	83.68	75.11
Attention U-Net [26]	80%	20%	85.00	73.00
BUSSAM [29]	80%	20%	89.95	82.31
HA-Net [4]	80%	10%	**97.28**	**94.75**
Ours	0%	100%	70.77	54.76

A direct comparison of methods evaluated on BUSI is confounded by two factors: (1) different train/validation split ratios, where methods using 80% of data for training have a substantial advantage over our evaluation setting; and (2) the well-documented duplicate structure of the BUSI dataset. As discussed in Section 4.4, a large proportion of images in BUSI have near-identical copies, meaning that any method relying on a randomly constructed split risks placing duplicate images in both the training and validation sets simultaneously. Under such conditions, reported metrics may be substantially overestimated and not reflective of true generalization capability. We therefore treat BUSI results from methods using random splits with caution, and motivate the use of an independent held-out benchmark for reliable evaluation.

5.3 Cross-Dataset Generalization on BUS-UC

To provide a contamination-free and fair comparison, we retrained BUSSAM and HA-Net using their official codebases on their respective data splits, and subsequently evaluated all three models on the BUS-UC dataset, which was not used in any model's training pipeline. Results are reported in Table 2.

Table 2: Cross-dataset generalization results on the BUS-UC dataset. All models were trained using their official data splits and evaluated without any fine-tuning on BUS-UC. Higher values indicate better generalization.

Method	**DSC (%)**	**IoU (%)**
HA-Net [4] (2025)	85.05	76.69
BUSSAM [29] (2024)	88.49	80.34
Ours	**92.30**	**85.73**

As shown in Table 2, our method achieves the highest DSC of **92.30%** and IoU of **85.73%** on the BUS-UC dataset, outperforming both BUSSAM (DSC: 88.49%, IoU: 80.34%) and HA-Net (DSC: 85.05%, IoU: 76.69%), demonstrating superior cross-dataset generalization. It is worth noting that our model was trained on a substantially larger dataset than the compared baselines, which may also contribute to its stronger generalization performance.

Notably, the performance ordering on BUS-UC (Ours > BUSSAM > HA-Net) differs markedly from the rankings reported on BUSI. This discrepancy is consistent with the duplicate-induced overfitting risk described in Section 4.4: models that achieve high scores on randomly partitioned BUSI splits may be benefiting from near-identical training and validation samples rather than demonstrating genuine generalization. Evaluation on the independent BUS-UC benchmark provides a more reliable basis for comparison.

Among the baselines, BUSSAM exhibits the most consistent performance across both datasets. However, this should be interpreted in light of its architectural design: BUSSAM is built upon SAM [17], which requires a visual prompt, such as a bounding box or point, explicitly indicating the lesion location at inference time. The model therefore performs boundary delineation given a known region of interest, rather than joint localization and segmentation. Our method, by contrast, receives no spatial prior at inference time and must independently identify and delineate lesions from the raw image alone, representing a strictly harder and more clinically realistic evaluation setting.

These findings confirm that duplicate-aware or independent-benchmark evaluation is essential for meaningful comparison in breast ultrasound segmentation, and we recommend BUS-UC as the standard going forward.

5.4 Ablation Study

Table 3 incrementally validates each component on BUS-UC. The baseline (R1) stacks the grayscale image three times as input; replacing this with a complementary original/CLAHE/contrast stack and unfreezing the first patch-embedding layer of USFM (R2) yields the largest single gain in IoU (+1.5%) and cuts HD95 by 2.2 px. Textual guidance alone (R3) adds just a notable boundary improvement (HD95 17.6 px), while multi-organ supporting data alone (R4) unexpectedly *degrades* DSC to 91.23% vs. R2's 91.63%, suggesting cross-organ diversity introduces ambiguity that semantic prompts are necessary to resolve, their combination (R5) recovers and surpasses both at 91.99%. Finally, the weak bounding-box pipeline (R6) delivers the largest boundary gain, reducing HD95 from 16.5 to 15.2 px, and lifts the full model to **92.30%** DSC and **85.73%** IoU. It is interesting to note that precisely the weakly annotated images with no pixel-level tumor boundary annotations contribute to the boundary accuracy, thanks to our proposed weak branch architecture.

6 Conclusion

We presented a semi-supervised breast ultrasound segmentation framework that fuses a frozen USFM visual backbone with CLIP-based semantic guidance via per-scale cross-attention. A teacher–student EMA scheme with dynamic loss scheduling enables effective exploitation of large-scale bounding-box annotations alongside pixel-precise masks, reducing dependence on costly pixel-level labeling. Evaluated in a zero-shot cross-dataset setting, our method achieves a state-of-the-art DSC of 92.30% and IoU of 85.73% on BUS-UC, outperforming both BUSSAM and HA-Net under identical evaluation conditions.

As discussed in Section 4.4, the widespread use of randomly partitioned BUSI splits in the literature carries a significant risk of duplicate-induced overfitting, which may inflate reported Dice scores and obscure true generalization ability. We urge the community to adopt independent benchmarks such as BUS-UC for future comparisons and to account for dataset duplicates when constructing evaluation protocols.

A particularly promising direction for future work concerns the nature of the text prompts themselves: in this work, prompts are limited to generic organ-class descriptions due to the absence of clinically grounded annotations. Real diagnostic reports authored by radiologists, capturing lesion echogenicity, shape irregularity, and surrounding tissue characteristics, could serve as far richer supervisory signals. As no such paired image–report datasets exist at scale for breast ultrasound today, we view their curation and exploitation as an important open problem for the community.

References

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

[2] Mohammed Alotaibi, Abdulrhman Aljouie, Najd Alluhaidan, Wasem Qureshi, Hessa Almatar, Reema Alduhayan, Barrak Alsomaie, and Ahmed Almazroa. Breast cancer classification based on convolutional

Table 3: Ablation study on the BUS-UC test set. ✓ = component active. HD95 in px (↓); Dice and IoU in % (↑).

ID	Configuration	Strong breast	Ch. preproc.	Supporting	Text guid.	Weak pipeline	Dice (%)	IoU (%)	HD95↓
R1	Baseline	✓					90.773	83.169	20.644
R2	+ Ch. preprocessing	✓	✓				91.628	84.649	18.433
R3	+ Textual guidance	✓	✓		✓		91.710	84.680	17.648
R4	+ Supporting data	✓	✓	✓			91.228	83.962	17.155
R5	+ Supporting + textual	✓	✓	✓	✓		91.988	85.236	16.515
R6	+ **Weak branch (Full pipeline)**	✓	✓	✓	✓	✓	**92.300**	**85.727**	**15.243**

neural network and image fusion approaches using ultrasound images. *Heliyon*, 9(11), 2023.

[3] Ali Abbasian Ardakani, Afshin Mohammadi, Mohammad Mirza-Aghazadeh-Attari, and U Rajendra Acharya. An open-access breast lesion ultrasound image database: Applicable in artificial intelligence studies. *Computers in Biology and Medicine*, 152:106438, 2023.

[4] Muhammad Azeem Aslam, Asim Naveed, Nisar Ahmed, and Zhang Ke. A hybrid attention network for accurate breast tumor segmentation in ultrasound images. *Scientific Reports*, 15(1):39633, 2025.

[5] 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.

[6] M Jorge Cardoso, Wenqi Li, Richard Brown, Nic Ma, Eric Kerfoot, Yiheng Wang, Benjamin Murrey, Andriy Myronenko, Can Zhao, Dong Yang, et al. Monai: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701*, 2022.

[7] Behnaz Gheflati and Hassan Rivaz. Vision transformers for classification of breast ultrasound images. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 480–483. IEEE, 2022.

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

[9] Haifan Gong, Guanqi Chen, Ranran Wang, Xiang Xie, Mingzhi Mao, Yizhou Yu, Fei Chen, and Guanbin Li. Multi-task learning for thyroid nodule segmentation with thyroid region prior. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 257–261, 2021.

[10] Haifan Gong, Jiaxin Chen, Guanqi Chen, Haofeng Li, Guanbin Li, and Fei Chen. Thyroid region prior guided attention for ultrasound segmentation of thyroid nodules. *Computers in biology and medicine*, 155:106389, 2023.

[11] Kelei He, Chen Gan, Zhuoyuan Li, Islem Rekik, Zihao Yin, Wen Ji, Yang Gao, Qian Wang, Junfeng Zhang, and Dinggang Shen. Transformers in medical image analysis. *Intelligent Medicine*, 3(1):59–78, 2023.

[12] Weijie He, Runyuan Bao, Yiru Cang, Jianjun Wei, Yang Zhang, and Jiacheng Hu. Axial attention transformer networks: A new frontier in breast cancer detection. *arXiv preprint arXiv:2409.12347*, 2024.

[13] Jin Huang, Yazhao Mao, Jingwen Deng, Zhaoyi Ye, Yimin Zhang, Jingwen Zhang, Lan Dong, Hui Shen, Jinxuan Hou, Yu Xu, et al. Emganet: Edge-aware multi-scale group-mix attention network for breast cancer ultrasound image segmentation. *IEEE Journal of Biomedical and Health Informatics*, 2025.

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

[15] Ahmed Iqbal and Muhammad Sharif. Memory-efficient transformer network with feature fusion for breast tumor segmentation and classification task. *Engineering Applications of Artificial Intelligence*, 127:107292, 2024.

[16] Jing Jiao, Jin Zhou, Xiaokang Li, Menghua Xia, Yi Huang, Lihong Huang, Na Wang, Xiaofan Zhang, Shichong Zhou, Yuanyuan Wang, et al. Usfm: A universal ultrasound foundation model generalized to tasks and organs towards label efficient image analysis. *Medical image analysis*, 96:103202, 2024.

[17] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

[18] Zhi Lin, Junhao Lin, Lei Zhu, Huazhu Fu, Jing Qin, and Liansheng Wang. A new dataset and a baseline model for breast lesion detection in ultrasound videos. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 614–623. Springer, 2022.

[19] Jie Liu, Yixiao Zhang, Jie-Neng Chen, Junfei Xiao, Yongyi Lu, Bennett A Landman, Yixuan Yuan, Alan Yuille, Yucheng Tang, and Zongwei Zhou. Clip-driven universal model for organ segmentation and tumor detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 21152–21164, 2023.

[20] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature communications*, 15(1):654, 2024.

[21] Yuhao Mo, Chu Han, Yu Liu, Min Liu, Zhenwei Shi, Jiatai Lin, Bingchao Zhao, Chunwang Huang, Bingjiang Qiu, Yanfen Cui, et al. 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.

[22] 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.

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[24] 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.

[25] Xiaoyan Shen, Liangyu Wang, Yu Zhao, Ruibo Liu, Wei Qian, and He Ma. Dilated transformer: residual axial attention for breast ultrasound image segmentation. *Quantitative Imaging in Medicine and Surgery*, 12(9):4512, 2022.

[26] Adel Sulaiman, Vatsala Anand, Sheifali Gupta, Adel Rajab, Hani Alshahrani, Mana Saleh Al Reshan, Asadullah Shaikh, and Mohammed Hamdi. Attention based unet model for breast cancer segmentation using busi dataset. *Scientific Reports*, 14(1):22422, 2024.

[27] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

[28] Zhi Tian, Chunhua Shen, Xinlong Wang, and Hao Chen. Boxinst: High-performance instance segmentation with box annotations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5443–5452, 2021.

[29] Z Tu, L Gu, X Wang, B Jiang, and A Provincial. Ultrasound sam adapter: Adapting sam for breast lesion segmentation in ultrasound images. arxiv 2024. *arXiv preprint arXiv:2404.14837*.

[30] 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.

[31] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434, 2018.

[32] Yiming Xu, Bowen Zheng, Xiaohong Liu, Tao Wu, Jinxiu Ju, Shijie Wang, Yufan Lian, Hongjun Zhang, Tong Liang, Ye Sang, et al. Improving artificial intelligence pipeline for liver malignancy diagnosis using ultrasound images and video frames. *Briefings in Bioinformatics*, 24(1):bbac569, 2023.

[33] Haonan Yang and Dapeng Yang. Cswin-pnet: A cnn-swin transformer combined pyramid network for breast lesion segmentation in ultrasound images. *Expert Systems with Applications*, 213:119024, 2023.

[34] Qi Zhao, Shuchang Lyu, Wenpei Bai, Linghan Cai, Binghao Liu, Guangliang Cheng, Meijing Wu, Xiubo Sang, Min Yang, and Lijiang Chen. Mmotu: A multi-modality ovarian tumor ultrasound image dataset for unsupervised cross-domain semantic segmentation. *arXiv preprint arXiv:2207.06799*, 2022.

[35] Chengzhang Zhu, Xian Chai, Yalong Xiao, Xu Liu, Renmao Zhang, Zhangzheng Yang, and Zhiyuan Wang. Swin-net: A swin-transformer-based network combing with multi-scale features for segmentation of breast tumor ultrasound images. *Diagnostics*, 14(3):269, 2024.