CSC-PA: Cross-image Semantic Correlation via Prototype Attentions for Single-network Semi-supervised Breast Tumor Segmentation

Zhenhui Ding[1], Guilian Chen[1], Qin Zhang[1], Huisi Wu[1]*, Jing Qin[2]
[1] College of Computer Science and Software Engineering, Shenzhen University
[2] Centre for Smart Health, School of Nursing, The Hong Kong Polytechnic University
2300271025@email.szu.edu.cn, hswu@szu.edu.cn

Abstract

Accurate automatic breast ultrasound (BUS) image segmentation is essential for early breast cancer screening and diagnosis. However, it remains challenging owing to (1) breast lesions of various scale and shape, (2) ambiguous boundaries caused by speckle noise and artifacts, and (3) the scarcity of high-quality annotations. Most existing semi-supervised methods employ the mean-teacher architecture, which merely learns semantic information within a single image and heavily relies on the performance of the teacher model. Therefore, we present a novel cross-image semantic correlation semi-supervised framework, named CSC-PA*, to improve the performance of BUS image segmentation. CSC-PA is trained based on a single network, which integrates a foreground prototype attention (FPA) and an edge prototype attention (EPA). Specifically, FPA transfers complementary foreground information for more stable and complete lesion segmentation. On the other hand, EPA enhances edge features of lesions by using edge prototype, where an adaptive edge container is proposed to store global edge features and generate the edge prototype. Additionally, we introduce a pixel affinity loss (PAL) to exploit previously ignored contextual correlation in supervision, which further improves performance on edges. Extensive experiments on two benchmark BUS datasets demonstrate that our model outperforms other state-of-the-art methods under different partition protocols. Codes are available at* https://github.com/shdkdh/CSC-PA.

1. Introduction

Breast cancer is one of the most common malignant tumors in women, threatening their health and life [6, 22]. Early detection and diagnosis play a crucial role in improving patients' quality of life, extending their lifespan, and reducing the mortality rate [8, 32]. Breast ultrasound (BUS) has been widely used in clinical practice owing to its radiation-free, non-invasive, cost-effective, and real-time nature [25]. However, conventional disease screening and detection of BUS images heavily relies on manual delineation, which is labor-intensive, time-consuming, and sensitive to the experience of radiologists. To overcome above limitations, some computer-aided diagnostic (CAD) systems equipped with automatic segmentation models have been developed to enhance the efficiency and accuracy of breast cancer screening. However, automatic BUS image segmentation remains a challenging task owing to (1) limited annotations, (2) the large variations of targets in terms of scale and shape (Fig. 1 (a)-(c)), and (3) the ambiguous boundaries between the lesion and the background caused by low contrast, speckle noise, and artifacts in BUS images (Fig. 1 (d)-(e)).

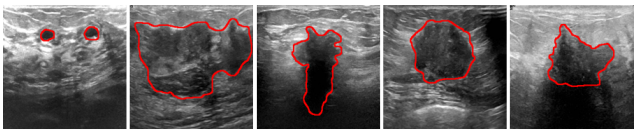

Figure 1. Main challenges of BUS image segmentation: (a)-(c) large variation in the scale and shape of breast tumors; (d)-(e) ambiguous boundaries. Red contours outline ground truth.

To tackle these challenges, early studies were devoted to exploring fully-supervised paradigms [4, 11, 20, 34], which have achieved remarkable performance. However, these models are heavily relies on extensive manual annotations. To alleviate the scarcity of labels, several researchers have begun to explore semi-supervised paradigms, attempting to learn the semantic information contained in massive unlabeled data. Mainstream semi-supervised methods can be divided into two categories: consistency regularization and pseudo-labeling. Consistency regularization enforces alignment between the original and perturbed outputs, including images, features, and network perturbations, and thus reduces the reliance on manual labels [5, 15, 17, 36, 41]. On the other hand, pseudo-labeling improves the model per-

*Corresponding Author

formance by generating extra training samples [2, 3, 12, 18, 28, 40]. Both paradigms are primarily implemented on the classic mean-teacher (MT) framework [27]. However, MT framework merely learns semantic correlations within a single image and heavily relies on the performance of the teacher model, which makes the teacher network susceptible to noise in individual images, particularly in scenarios with scarce annotations. Additionally, as, in the training procedure, student model may learn inaccurate representations from the teacher model and feeds these errors back to the teacher network, the framework is vulnerable to challenging regions between lesions and background, leading to overall performance degradation. Wu et al. [30] improve the utilization of unlabeled data by identifying the most similar labeled image and employing contrastive learning to correct pseudo-labels based on semantic correlations across images. However, this method largely ignores the potential positive impact of cross-image information on labeled data.

In this paper, we propose a novel cross-image semantic correlation semi-supervised framework for BUS image segmentation, which integrates two prototype attentions into a single network. Specifically, we propose a foreground prototype attention (FPA) that leverages channel prototypes and an attention mechanism to transfer complementary foreground features between labeled and unlabeled images. FPA is able to capture a broader feature distribution and mitigate the negative impact of noise in individual images by mining semantic correlation across images, resulting in more stable and complete segmentation outcomes. To improve segmentation at blurred boundaries, we further propose an edge prototype attention (EPA), which uses edge prototype to enhance edge representation learning. In EPA, we design a novel adaptive edge container for storing global edge features, diverging from traditional memory banks that update in a first-in-first-out manner. Furthermore, we introduce a pixel affinity loss (PAL) to investigate the previously ignored contextual correlations in ground truth (GT) and pseudo-labels, by enforcing the consistency of affinity matrices derived from predictions and corresponding supervision. PAL is beneficial to improve the model's precise identification of lesion edges. Extensive experiments on BUSI [1] and UDIAT [39] datasets show that our method consistently outperforms other state-of-the-art models under the same experimental settings, demonstrating the effectiveness of our model. Our contributions are summarized as follows:

- We present a novel cross-image semantic correlation semi-supervised framework, named *CSC-PA*, for BUS image segmentation, which integrates two proposed prototype attentions (FPA and EPA) to capture the semantic correlations across images and enhance edge representation learning, respectively.
- We propose the PAL to emphasize neglected contextual relations in supervision by constraining the consistency of affinity matrices generated from predictions and corresponding supervision, acquiring improved lesion edges.
- Our method achieves state-of-the-art (SOTA) performance on two benchmark BUS datasets, surpassing other competitors under different partition protocols.

2. Related Work

2.1. Breast Ultrasound Image Segmentation

Recently, U-shaped networks [13, 19, 26, 43] incorporate multi-scale processing and skip connections to enhance segmentation performance. Subsequently, transformer-based methods [9, 16, 35, 44] further improve segmentation results by considering global semantic relations. MFMSNet [31] introduces a multi-scale interactive fusion module and a multi-frequency transformer block to refine lesion edges. Although the above fully-supervised methods demonstrate remarkable results, they heavily rely on a large amount of expensive annotations. To alleviate this issue, some semi-supervised methods for BUS image segmentation have been proposed. DK-HRS [33] adopts a hybrid regularization strategy, which integrates a virtual adversarial training module and a contrastive learning based on domain knowledge into the FixMatch [23], achieving superior performance. PH-Net [10] introduces an adaptive patch augmentation and a hard-patch contrastive learning to improve the feature learning in hard regions and guide the model's attention to challenging pixels in difficult areas, respectively.

2.2. Semi-supervised Semantic Segmentation

The primary challenge in semi-supervised semantic segmentation is to effectively exploit the abundant unlabeled data. In MT architecture, U^2PL [29] treats low-confidence pixels as negative samples for contrastive learning, maximizing the utilization of unlabeled images. AugSeg [42] employs an intensity-based augmentation to enhance adaptability in semi-supervised scenarios. PS-MT [15] introduces a new auxiliary teacher on MT framework, and proposes a confidence-weighted cross-entropy loss to address the issue of inaccurate predictions in consistency learning. UniMatch [36] and CorrMatch [24] are both single network, where the former proposes a dual-stream perturbation technique to exploit a broader perturbation space, and the latter designs two label propagation based on correlation maps to maximize the utilization of unlabeled data.

2.3. Affinity Matrix

The affinity matrix captures the semantic similarity between pixels, reflecting the probability that two points belong to the same category [14]. Generally, a higher similarity between two points corresponds to an affinity closer to 1, while a lower similarity yields an affinity closer to 0. In connectivity analysis, this matrix is known as the connectiv-

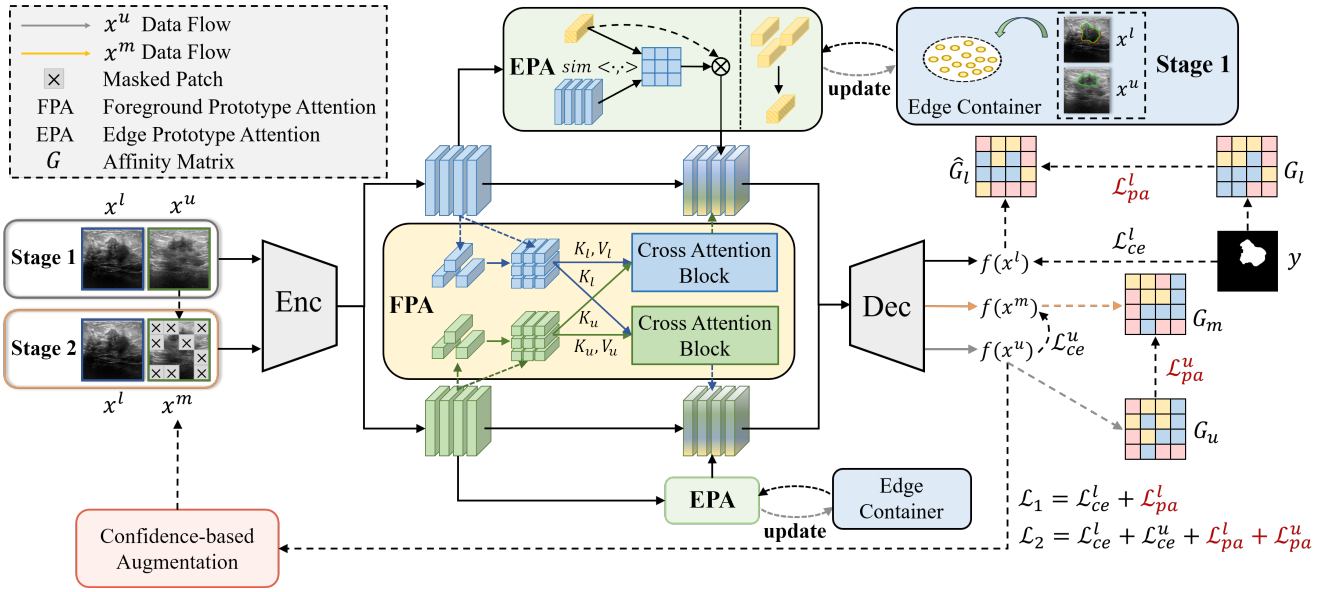

Figure 2. Overview of our CSC-PA with single network training framework. We introduce a foreground prototype attention (FPA) to extract semantic correlations across different images for more stable and accurate segmentation. Besides, we propose an edge prototype attention (EPA) to amplify edge features and enhance edge representation learning, along with a novel edge container to store global edge features. In stage 1, the input unlabeled image is the original image x^u, while in stage 2, the input is x^m obtained by applying confidence-based augmentation to x^u. The corresponding prediction $f(x^u)$ is used to supervise $f(x^m)$. In addition, we propose a pixel affinity loss (PAL) to further extend the attention on more edge areas by mining the contextual correlations of ground truth y and pseudo-label $f(x^u)$.

ity matrix, which facilitates the examination of connectivity between pixels and their neighbors [37, 38]. In this paper, we employ the affinity matrix to model the semantic correlations between pixels and their surrounding pixels, which further enhances edge segmentation results.

3. Method

3.1. Overview

The presence of speckle noise and artifacts in BUS images, combined with the scarcity of labeled data, poses significant challenges for accurate lesion segmentation. We propose a novel cross-image semantic correlation semi-supervised framework, named CSC-PA, to improve segmentation performance for BUS task. As illustrated in Fig. 2, in contrast to traditional MT framework, we integrate a foreground prototype attention (FPA) and an edge prototype attention (EPA) into a single network architecture. Specifically, FPA utilizes channel prototypes and an attention mechanism to facilitate the transfer of foreground information, resulting in more stable and comprehensive segmentation outcomes. EPA employs an edge prototype to amplify edge features and enhance edge representation learning, which is beneficial to refine edge segmentation. In EPA, we design a novel adaptive edge container to store global edge features and generate edge prototype. Furthermore, we propose a pixel affinity loss (PAL) to mine previously overlooked contextual correlations in supervision. By modeling contextual relations through affinity matrix and imposing KL divergence as a constraint, the model effectively captures semantic correlations among pixels within a single image, thereby improving model performance, especially in lesion edges.

3.2. Foreground Prototype Attention

Most semi-supervised approaches ignore cross-image semantic correlations, making models susceptible to noise, particularly in limited annotation scenarios. Therefore, we propose FPA, which leverages channel prototypes and an attention mechanism to transfer complementary foreground information between labeled and unlabeled images. FPA not only enables the model to capture rich cross-image semantic information and a broader feature distribution, but also mitigates the influence of noise and artifacts from individual images, addressing some ambiguous or erroneous predictions. Meanwhile, FPA allows unlabeled images to indirectly share labeled annotations, facilitating the semantic extraction from unlabeled images.

As shown in Fig. 3, we visualize the implementation process of FPA. Inspired by [21], we observe that the activation region of each channel feature tends to encode the salience of the image scene categories. Building on this, we derive channel prototypes from the feature $F \in \mathbb{R}^{B\times C\times HW}$ extracted by the encoder, where B and C represent the batch size and the number of channels, respectively. H and W are

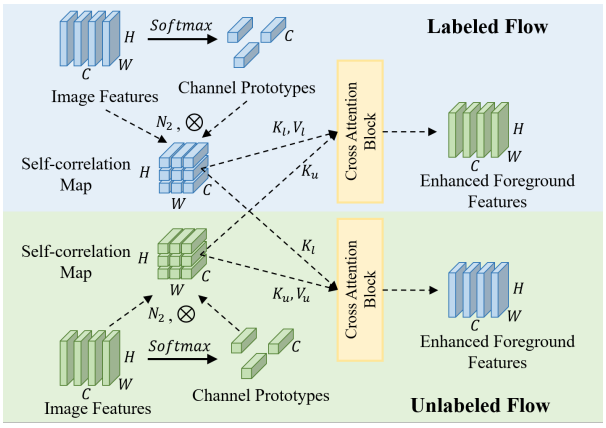

Figure 3. Illustration of foreground prototype attention (FPA).

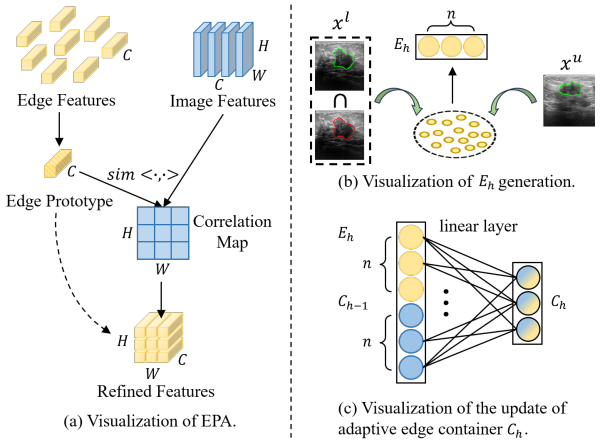

Figure 4. Illustration of edge prototype attention (EPA). (a) represents the working schema of EPA. (b)-(c) depict the generation of E_h and the update of adaptive edge container C_h, respectively.

the length and width of the feature map, respectively. To emphasize salient features while suppressing insignificant ones, we first apply a softmax operation along the channel dimension of each feature map, which quantify the contribution of each channel to each pixel. Then the channel features are weighted according to the $\mathrm{Softmax}(F)$, highlighting lesion-related features in each channel:

$$P_c = F \otimes \mathrm{Softmax}(F)^T, \tag{1}$$

where $\otimes$ is matrix multiplication. For channel prototypes $P_c \in \mathbb{R}^{B\times C\times C}$, each prototype represents salient lesion features encoded within channel. To integrate the salient feature from generated channel prototypes P_c with image features F, we compute the self-correlation map $\Phi_f \in \mathbb{R}^{B\times C\times HW}$ for labeled and unlabeled images as follows:

$$\Phi_f = \mathcal{N}_2(P_c)^T \otimes \mathcal{N}_2(F), \tag{2}$$

where $\mathcal{N}_2$ is L_2 normalization. Φ_f indicates the similarity between image features and the centers of lesion prototypes, which facilitates the identification of lesion region. Since each batch contains an equal number of labeled and unlabeled samples, we have $\Phi_f = \mathrm{Concat}(\Phi_f^l, \Phi_f^u)$, where Φ_f^l and Φ_f^u share the same dimension in the first axis.

Subsequently, we adopt an attention mechanism to extract complementary foreground features for transfer between labeled and unlabeled images. For labeled and unlabeled data, the enhanced foreground features F_f^l and F_f^u are calculated as follows:

$$K_l = \mathbf{L}(\Phi_f^l),\ V_l = \mathbf{L}(\Phi_f^l), \tag{3}$$

$$K_u = \mathbf{L}(\Phi_f^u),\ V_u = \mathbf{L}(\Phi_f^u), \tag{4}$$

$$F_f^l = \mathrm{Softmax}(\frac{K_l K_u^T}{\sqrt{C}})V_l, \tag{5}$$

$$F_f^u = \mathrm{Softmax}(\frac{K_u K_l^T}{\sqrt{C}})V_u, \tag{6}$$

where $\mathbf{L}$ denotes the linear layer, and the enhanced foreground features of a batch are represented as $F_f = \mathrm{Concat}(F_f^l, F_f^u)$. Then, we incorporate F_f into the image features F, which improves the distinction between the lesion and background regions, facilitating more stable and comprehensive segmentation of lesions.

Additionally, for unlabeled data, we design a confidence-based augmentation strategy to generate masked image x^m, which selectively masks simple regions of x^u based on the prediction $f(x^u)$. By inferring overall prediction from complex regions and limited simple regions, the model can extract additional mutual information from the challenging areas, thereby improving segmentation performance. Further details are provided in the appendix.

3.3. Edge Prototype Attention

Although FPA transfers semantic information between images, the model struggles to achieve accurate segmentation in challenging areas, particularly at lesion edges. Therefore, we propose EPA to obtain refined edge segmentation results by enhancing edge representation learning. Specifically, we design an adaptive edge container to store global edge features, from which an edge prototype is generated by averaging all features within the container. Then, we utilize cosine similarity between the obtained edge prototype and image features to amplify edge features of input images.

In Fig. 4, we present the main operations of EPA. First, we design an adaptive edge container to store global edge features, which dynamically filters and updates these features during training. For container initialization, we randomly select n correctly predicted edge features from the labeled data at the end of pre-training to serve as the initial edge container C_h, where h denotes the current iteration. In stage 1 of subsequent iteration, as shown in Fig. 4 (b),

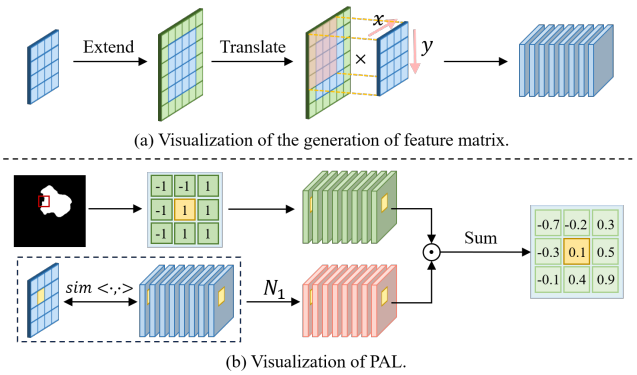

Figure 5. Illustration of pixel affinity loss (PAL). (a) depicts the expansion and translation involved in feature matrix generation. (b) represents the calculation process of PAL.

we put edge features from labeled image predictions that align with GT, and features of high-confidence edge pixels in unlabeled images into a set. Then, we randomly select n features from the set as current edge set E_h. In Fig. 4 (c), different from memory bank (MB), which stores features in a first-in-first-out manner, our container C_h adaptively filters and updates edge features through a linear layer:

$$C_h = \mathbf{L}(\text{Concat}(E_h, C_{h-1})). \tag{7}$$

Due to the random selection and mapping mechanism of linear layer, our edge container effectively captures diverse edge features and generates a representative edge prototype, which is conducive to obtaining accurate edge segmentation results. Subsequently, we compute the edge prototype $P_e \in \mathbb{R}^{C\times1\times1}$ by averaging all features contained in C_h:

$$P_e = \frac{1}{|C_h|}\sum_{x_i\in C_h} x_i, \tag{8}$$

where $|C_h|$ is the number of edge features stored in C_h.

As shown in Fig. 4 (a), we leverage the cosine similarity between the edge prototype and image features to enhance edge details. Specifically, we first calculate the correlation map Φ_e between the prototype and image features:

$$\Phi_e = \text{Softmax}(sim\,\langle P_e, F\rangle), \tag{9}$$

where $sim\,\langle\cdot,\cdot\rangle$ is the cosine similarity. The obtained correlation map Φ_e represents the probability of each pixel belonging to the lesion edges. The softmax function is employed to amplify the probability values of prominent edge pixels within the correlation map, which enhances the distinction between edge regions and others. Then, we multiply Φ_e by the edge prototype P_e to obtain the enhanced edge features F_e, calculated as follows:

$$F_e = \Phi_e \odot P_e, \tag{10}$$

where $\odot$ is dot product. The final image features F' fed into the decoder are:

$$F' = F \oplus F_f \oplus F_e, \tag{11}$$

where $\oplus$ represents matrix addition.

3.4. Pixel Affinity Loss

Although cross-entropy loss ℓ_{ce} is widely used in semi-supervised semantic segmentation, it primarily enforces pixel-wise consistency, while neglecting the contextual correlations between pixels within the supervision. Therefore, we propose PAL to constrain the consistency of contextual relations between predictions and corresponding GT or pseudo-labels. This loss is conducive to fully exploiting the supervisory signals, leading to more accurate segmentation results, especially in lesion edges. As shown in Fig. 5 (a), we first employ expansion and shift operations to generate the feature matrix $A \in \mathbb{R}^{8\times HW}$, which contains the features of the eight surrounding pixels for each pixel. Similarly, we construct the connection matrix M from annotations or pseudo-labels, where foreground pixels are mapped to 1 and background pixels to -1. Subsequently, we calculate the cosine similarity between each pixel feature and its corresponding values in A to quantify the affinity similarity, followed by the application of $\mathcal{N}_1$ regularization. The resulting values are then multiplied by the connection matrix M, and summed along the channel dimension:

$$G = \sum \mathcal{N}_1(sim\,\langle F, A\rangle) \odot M, \tag{12}$$

where G represents the affinity matrix, which models the correlations of each pixel in relation to its surrounding pixels. For GT, surrounding pixels contribute equally in correlation modeling. Considering the potential errors in pseudo-labels, pixels with confidence below threshold τ are excluded from correlation modeling. The labeled pixel affinity loss $\mathcal{L}^l_{pa}$ and unlabeled pixel affinity loss $\mathcal{L}^u_{pa}$ are defined as:

$$\mathcal{L}^l_{pa} = \text{KL}(\hat{G}_l, G_l), \tag{13}$$
$$\mathcal{L}^u_{pa} = \text{KL}(G_m, G_u), \tag{14}$$

where KL denotes the Kullback-Leibler divergence, utilized to minimize the discrepancy in affinity matrices derived from predictions and corresponding supervision. The $\hat{G}_l$ and G_m are obtained from predictions of labeled image x^l and masked unlabeled image x^m, while G_l and G_u are derived from GT y and pseudo-label $f(x^u)$, respectively.

3.5. Loss Functions

We employ ℓ_{ce} to constrain the pixel-wise consistency between the predictions and corresponding ground truth (GT)

Method	Year	BUSI						UDIAT					
		1/2 (258)		1/4 (129)		1/8 (64)		1/2 (66)		1/4 (33)		1/8 (16)	
		Dice±std(%)	IoU±std(%)	Dice±std(%)	IoU±std(%)	Dice±std(%)	IoU±std(%)	Dice±std(%)	IoU±std(%)	Dice±std(%)	IoU±std(%)	Dice±std(%)	IoU±std(%)
U^{2}PL	2022	75.86±1.71	66.78±1.49	73.61±2.26	64.37±2.80	71.19±3.53	61.89±2.93	80.22±3.15	70.40±3.84	77.51±2.87	67.64±3.80	75.29±2.35	65.26±3.06
AugSeg	2023	76.62±1.94	67.93±1.77	75.54±2.49	66.62±2.52	72.25±1.98	63.12±2.16	81.39±3.72	71.77±4.36	79.48±4.37	69.81±4.25	77.24±2.81	67.63±3.07
ABD	2024	77.14±2.69	68.80±2.47	75.96±2.81	67.36±3.09	72.70±2.24	63.44±2.57	81.87±2.02	72.28±2.97	80.40±2.24	70.73±2.11	77.63±4.94	68.04±5.31
CISC-R	2023	76.91±1.99	68.62±2.25	75.78±2.85	67.02±3.04	72.59±3.59	63.24±3.02	81.53±3.28	71.98±3.47	79.82±4.00	70.19±3.57	77.57±3.46	67.92±3.85
PS-MT	2022	75.67±1.62	66.35±2.10	74.22±1.23	65.12±1.55	71.35±2.53	62.09±3.36	80.54±3.32	70.76±4.24	78.26±2.48	68.50±3.28	76.49±3.09	66.63±3.42
UniMatch	2023	76.30±2.01	67.49±2.29	75.08±2.27	66.04±2.70	71.97±2.80	62.70±2.48	80.82±3.50	71.15±4.71	78.81±3.03	69.17±3.49	76.80±4.84	67.01±5.18
CorrMatch	2024	76.96±2.20	68.57±2.55	75.83±2.69	67.10±2.61	72.57±3.01	63.17±3.34	81.65±3.07	72.04±3.59	80.09±2.76	70.36±2.93	77.44±4.28	67.82±4.21
MFMSNet	2024	75.22±1.97	65.92±1.74	72.77±2.80	62.96±2.63	69.02±2.35	59.41±2.27	78.50±2.39	68.82±2.55	76.23±2.90	66.05±3.03	73.15±3.63	62.94±3.20
DK-HRS	2024	76.79±2.27	68.33±1.98	75.63±2.26	66.86±2.14	72.38±2.17	63.06±2.59	81.47±3.53	71.86±4.10	79.74±3.88	70.01±3.59	77.38±3.09	67.75±3.26
PH-Net	2024	77.51±2.11	69.15±2.36	76.21±2.42	67.67±2.66	73.06±2.92	63.82±2.74	82.25±2.41	72.60±3.38	80.52±2.75	70.87±3.31	77.93±4.58	68.46±4.73
SupOnly	-	74.90±2.19	65.59±2.05	71.49±2.27	61.52±2.15	66.90±2.20	56.01±2.26	77.93±3.11	68.18±3.59	74.78±2.44	64.39±2.76	71.17±2.51	60.77±3.07
Ours	-	**78.24±1.81**	**69.63±1.89**	**76.93±2.28**	**68.20±2.22**	**74.10±1.95**	**64.58±1.82**	**82.71±2.57**	**73.14±2.70**	**81.26±1.75**	**71.65±1.89**	**78.70±4.02**	**69.07±4.72**

Table 1. Quantitative comparisons using different SOTA methods on the BUSI and UDIAT test sets.

or pseudo-labels. The pixel-wise supervision of the labeled and unlabeled images are defined as:

$$\mathcal{L}_{ce}^{l} = \ell_{ce}(f(x^{l}), y), \tag{15}$$

$$\mathcal{L}_{ce}^{u} = \ell_{ce}(f(x^{m}), f(x^{u})), \tag{16}$$

where $f(x^l)$, $f(x^u)$ and $f(x^m)$ represent the classification probabilities of the labeled image, unlabeled image and masked unlabeled image, respectively. y is the GT. The total losses for stage 1 and stage 2, denoted as $\mathcal{L}_1$ and $\mathcal{L}_2$, are formulated as follows:

$$\mathcal{L}_1 = \mathcal{L}_{ce}^{l} + \mathcal{L}_{pa}^{l}, \tag{17}$$

$$\mathcal{L}_2 = \mathcal{L}_{ce}^{l} + \mathcal{L}_{ce}^{u} + \mathcal{L}_{pa}^{l} + \mathcal{L}_{pa}^{u}. \tag{18}$$

4. Experiments

4.1. Experimental Setup

Datasets and Evaluation Metrics. We validate the effectiveness of our method on two famous BUS datasets.

BUSI [1] dataset comprises 780 images (133 normal, 437 benign and 210 malignant) with an average resolution of 500 × 500. These images were collected from 600 females aged 25 to 75 at Baheya Hospital in Egypt. In this paper, we conduct experiments on 647 abnormal images.

UDIAT [39] dataset is a relatively small dataset, which only contains 163 images (110 benign and 53 malignant) with an average resolution of 760 × 570. These images were collected from the UDIAT Diagnostic Centre in Spain.

We employ 5-fold cross-validation to evaluate our method. The training sets are further divided into labeled and unlabeled subsets following 1/2, 1/4 and 1/8 protocols. Segmentation performance is measured using the dice coefficient (Dice) and intersection over union (IoU).

Implementation Details. Our method is implemented using the PyTorch framework on a single NVIDIA GeForce RTX 3090 with 24 GB memory. Following previous semi-supervised work, we employ ResNet-50 pre-trained on ImageNet as the backbone and utilize Deeplabv3+ as the decoder. Our model is trained using a stochastic gradient descent (SGD) optimizer with momentum of 0.9 and weight decay of 0.0005. The learning rate is initialized to 0.003 and decayed using a polynomial strategy, $lr = lr_{in} \times (1 - \frac{t}{T})^{0.9}$, following [10]. At each training epoch, we equally sample labeled and unlabeled data, cropping input images to 513 × 513. The model is trained for a total of 200 epochs, with the first 10 epochs used for pre-training, and the batch size set to 8. n and τ are set to 256 and 0.8, respectively. Additionally, we randomly generate weak data augmentations for both labeled and unlabeled images, including random cropping, random scaling and random horizontal flipping.

4.2. Comparison with State-of-the-art Methods

We compare our method with ten SOTA competitors on the BUSI and UDIAT datasets under 1/2, 1/4 and 1/8 partition protocols. U^{2}PL [29], AugSeg [42] and ABD [7] are implemented on the MT framework. CISC-R [30] uses contrasive learning for cross-image modeling. PS-MT [15] introduces an auxiliary teacher on the MT architecture, while UniMatch [36] and CorrMatch [24] are single network. In addition, MFMSNet [31] is a fully-supervised method for BUS images. DK-HRS [33] and PH-Net [10] are semi-supervised BUS image segmentation methods. All methods are evaluated under the same experimental settings.

In Tab. 1, we compare our method with above SOTA approaches on the BUSI and UDIAT datasets. The results indicate that our CSC-PA obtains new SOTA performance across all partitioning protocols. We also observe that the performance improvements become more noticeable when fewer labels are available. Under the 1/2, 1/4 and 1/8 partition protocols of the UDIAT dataset, our method surpasses the baseline (SupOnly) by improvements of +4.96%, +7.26% and +8.30% in IoU, respectively. Notably, in the BUSI dataset, our method achieves a higher Dice of 76.93%

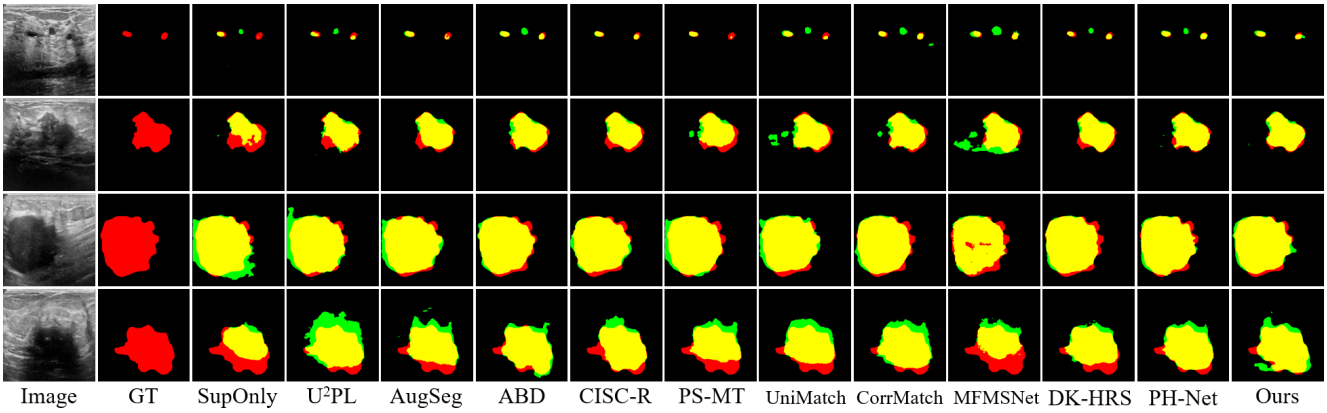

Figure 6. Visual comparison with different SOTA methods on two benchmark BUS test sets on 1/4 partitioning protocol. Red, green and yellow represent the ground truth, prediction and their overlapping regions, respectively.

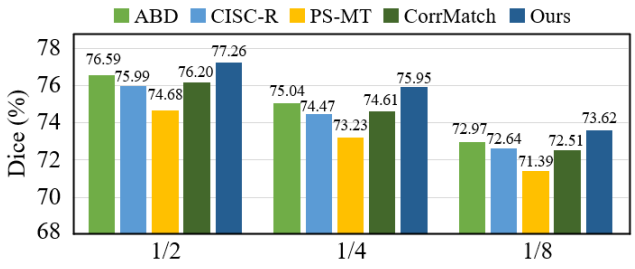

Figure 7. Comparison on unlabeled data of the BUSI dataset.

Ours			BUSI (129)		UDIAT (33)	
FPA	EPA	PAL	Dice±std(%)	IoU±std(%)	Dice±std(%)	IoU±std(%)
			71.49±2.27	61.52±2.15	74.78±2.44	64.39±2.76
✓			76.20±2.73	67.47±2.62	78.79±3.35	69.26±3.66
✓	✓		76.48±2.75	67.66±2.48	80.29±1.68	70.35±1.80
		✓	74.19±3.09	64.65±3.04	75.91±3.96	65.81±4.86
✓	✓	✓	**76.93±2.28**	**68.20±2.22**	**81.26±1.75**	**71.65±1.89**

Table 2. Ablation studies on CSC-PA. FPA: foreground prototype attention. EPA: edge prototype attention. PAL: pixel affinity loss.

under the 1/4 split protocol, while AugSeg obtains a Dice of 76.62% under the 1/2 split protocol. The results indicate that our method has potential in label-scare scenarios. We attribute these improvements to the two proposed prototype attentions (FPA and EPA) and the PAL.

Furthermore, we present a qualitative comparison of our method with SOTA methods in Fig. 6. The results indicate that our CSC-PA achieves stable and complete segmentation outcomes, even in cases with multiple lesions, significant scale variations and ambiguous boundaries. Additionally, to verify the effectiveness of our method in utilizing unlabeled images, we compare the Dice of the best performing methods from four different frameworks on the BUSI dataset: ABD, CISC-R, PS-MT, and CorrMatch. As shown

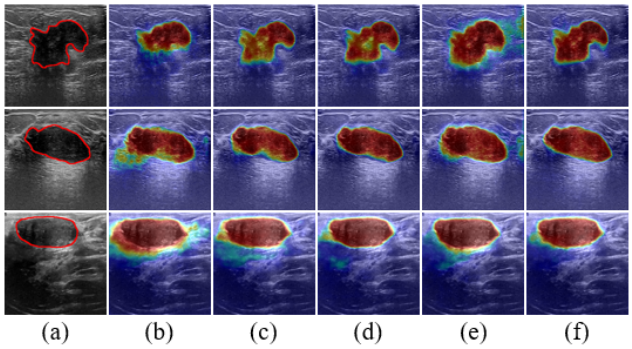

Figure 8. Visualization of module ablations. (a) Image and ground truth. (b) SupOnly. (c) w/FPA. (d) w/(FPA+EPA). (e) w/PAL. (f) w/(FPA+EPA+PAL). Red contours represent ground truth.

in Fig. 7, our method achieves the highest Dice across all partitions, demonstrating that CSC-PA effectively extracts more valuable information from unlabeled data for training.

4.3. Ablation Studies

We perform ablation studies on two BUS datasets using the 1/4 split strategy. The hyper-parameters ablations are presented in the appendix.

Ablation of components. As shown in Tab. 2 and Fig. 8, we conduct progressive ablation studies on each component of our CSC-PA. We adopt a single network as our baseline, which conducts supervised learning on labeled data. In Tab. 2, the integration of FPA leads to improvements in all metrics on both BUS datasets compared to the baseline. The results demonstrate that the proposed FPA effectively transfers complementary foreground knowledge. Subsequently, we integrate EPA with FPA, and the improvements indicate that the proposed EPA effectively supplements edge details, improving the model's performance. Furthermore, we incorporate PAL into the baseline and the single network

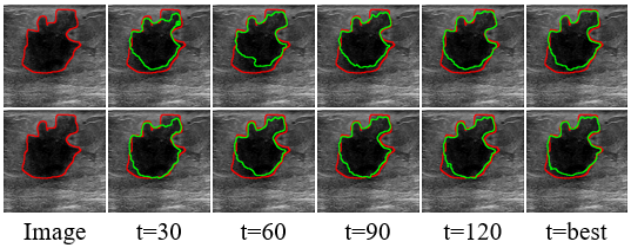

Figure 9. Segmentation results of memory bank (MB) and edge container across different epochs. The first row is MB and the second row is edge container. Red and green represent ground truth and predicted edges, respectively. t is the training epoch.

	Selection		Updating			UDIAT (33)	
	Top-k	Random	MB	CNN	Linear	Dice±std(%)	IoU±std(%)
I		✓	✓			80.28±1.55	70.41±1.78
II		✓		✓		79.95±1.80	69.92±1.39
III	✓				✓	80.45±2.22	70.31±2.16
IV		✓			✓	**81.26±1.75**	**71.65±1.89**

Table 3. Ablation studies of selection and updating methods. MB represents memory bank.

with above two prototype attentions, respectively. The results show that considering the consistency of affinity matrices derived from predictions and corresponding supervision significantly enhances model performance. In Fig. 8, with the progressive integration of components, the predictions exhibit more complete lesions and clearer boundaries, while demonstrating greater resilience to noise.

Effectiveness of adaptive edge container. To verify the effectiveness of the designed adaptive edge container in comparison to the traditional memory bank (MB), we conduct ablation studies on two edge storage approaches. As illustrated in Tab. 3 (I and IV), the proposed edge container exhibits superior performance, attributed to its adaptive edge feature learning mechanism, which effectively filters and updates edge features through a linear layer. Furthermore, Fig. 9 shows the edge segmentation results of the container and MB on test set during training, highlighting that the designed container achieves better lesion edge.

Impact of container update manner. During the edge container update process, we use a random selection to obtain current edge features E_h and employ a linear layer for adaptive filtering and updating the container C_h. For selection and updating methods, as illustrated in Tab. 3 (II, III and IV), we compare random selection with confidence-based top-k selection, as well as the linear layer with convolutional layer. The results show that our method achieves superior performance, due to the diverse edge features captured by random selection and the effective mapping mechanism of the linear layer. In Fig. 10, we visualize the t-SNE of above approaches, which indicates that our method yields a clearer separation between lesions and background.

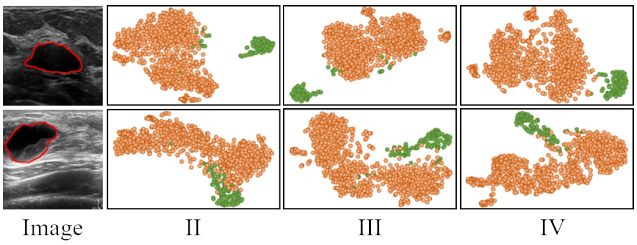

Figure 10. t-SNE visualization of features extracted by encoder. Orange represents the background, while green represents lesions.

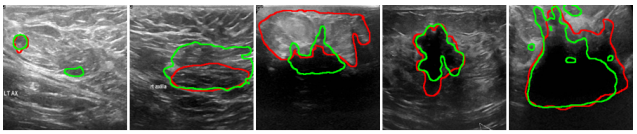

Figure 11. Failure cases. Red and green denote ground truth and predicted edges, respectively.

4.4. Discussions and Limitations

Although our method is tailored for BUS image segmentation, we believe that our framework has the potential for extension to other ultrasound tasks that present similar challenges. Moreover, our CSC-PA still has some limitations. In Fig. 11, extremely low contrast (a)-(b) and severe artifacts (c)-(e) may hinder the effectiveness of our approach.

5. Conclusion

In this paper, we present a novel cross-image semantic correlation semi-supervised framework, which integrates FPA and EPA into a single network architecture. To obtain stable and complete lesions in BUS images, FPA employs channel prototypes and an attention mechanism to transfer foreground features. EPA utilizes edge prototype to amplify edge features of input images, tackling the issue of blurred edges. Additionally, we propose PAL to mine contextual relations in supervision by enforcing the consistency of affinity matrices, further refining edges. Our method achieves SOTA performance on two BUS datasets under different partition protocols.

Acknowledgments

This work was supported partly by National Natural Science Foundation of China (Nos. 62273241 and 62206179), Natural Science Foundation of Guangdong Province, China (No. 2024A1515011946), Hong Kong RGC Theme-based Research Scheme (project no. T45-401/22-N), and Collaborative Research with World-leading Research Groups scheme of The Hong Kong Polytechnic University (project no. G-SACF).

References

[1] Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in brief*, 28:104863, 2020. 2, 6

[2] Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2020. 2

[3] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019. 2

[4] Michal Byra, Piotr Jarosik, Aleksandra Szubert, Michael Galperin, Haydee Ojeda-Fournier, Linda Olson, Mary O'Boyle, Christopher Comstock, and Michael Andre. Breast mass segmentation in ultrasound with selective kernel u-net convolutional neural network. *Biomedical Signal Processing and Control*, 61:102027, 2020. 1

[5] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2613–2622, 2021. 1

[6] Bhupender S Chhikara, Keykavous Parang, et al. Global cancer statistics 2022: the trends projection analysis. *Chemical Biology Letters*, 10(1):451–451, 2023. 1

[7] Hanyang Chi, Jian Pang, Bingfeng Zhang, and Weifeng Liu. Adaptive bidirectional displacement for semi-supervised medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4070–4080, 2024. 6

[8] Angela N Giaquinto, Hyuna Sung, Kimberly D Miller, Joan L Kramer, Lisa A Newman, Adair Minihan, Ahmedin Jemal, and Rebecca L Siegel. Breast cancer statistics, 2022. *CA: a cancer journal for clinicians*, 72(6):524–541, 2022. 1

[9] Qiqi He, Qiuju Yang, and Minghao Xie. Hctnet: A hybrid cnn-transformer network for breast ultrasound image segmentation. *Computers in Biology and Medicine*, 155: 106629, 2023. 2

[10] Siyao Jiang, Huisi Wu, Junyang Chen, Qin Zhang, and Jing Qin. Ph-net: Semi-supervised breast lesion segmentation via patch-wise hardness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11418–11427, 2024. 2, 6

[11] Huang Kai, Zhang Yu Feng, He Meng, Feng Yue Baoping, and Yao Rui Han. Ultrasound image segmentation of breast tumors based on swin-transformerv2. In *Proceedings of the 2022 10th International Conference on Information Technology: IoT and Smart City*, pages 106–111, 2022. 1

[12] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, page 896. Atlanta, 2013. 2

[13] Jia Liu, Jun Shao, Sen Xu, Zhiyong Tang, Weiquan Liu, Zeshuai Li, Tao Wang, and Xuesheng Bian. Asym-unet: An asymmetric u-shape network for breast lesions ultrasound images segmentation. *Biomedical Signal Processing and Control*, 99:106822, 2025. 2

[14] Sifei Liu, Shalini De Mello, Jinwei Gu, Guangyu Zhong, Ming-Hsuan Yang, and Jan Kautz. Learning affinity via spatial propagation networks. *Advances in Neural Information Processing Systems*, 30, 2017. 2

[15] Yuyuan Liu, Yu Tian, Yuanhong Chen, Fengbei Liu, Vasileios Belagiannis, and Gustavo Carneiro. Perturbed and strict mean teachers for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4258–4267, 2022. 1, 2, 6

[16] Zhenkun Lu, Chaoyin She, Wei Wang, and Qinghua Huang. Lm-net: A light-weight and multi-scale network for medical image segmentation. *Computers in Biology and Medicine*, 168:107717, 2024. 2

[17] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018. 1

[18] Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. *arXiv preprint arXiv:2101.06329*, 2021. 2

[19] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 2

[20] Bryar Shareef, Aleksandar Vakanski, Phoebe E Freer, and Min Xian. Estan: Enhanced small tumor-aware network for breast ultrasound image segmentation. In *Healthcare*, page 2262. MDPI, 2022. 1

[21] Changyong Shu, Yifan Liu, Jianfei Gao, Zheng Yan, and Chunhua Shen. Channel-wise knowledge distillation for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5311–5320, 2021. 3

[22] Rebecca L Siegel, Kimberly D Miller, Nikita Sandeep Wagle, and Ahmedin Jemal. Cancer statistics, 2023. *CA: a cancer journal for clinicians*, 73(1):17–48, 2023. 1

[23] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020. 2

[24] Boyuan Sun, Yuqi Yang, Le Zhang, Ming-Ming Cheng, and Qibin Hou. Corrmatch: Label propagation via correlation matching for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3097–3107, 2024. 2, 6

[25] Fenghe Tang, Lingtao Wang, Chunping Ning, Min Xian, and Jianrui Ding. Cmu-net: a strong convmixer-based medi-

cal ultrasound image segmentation network. In *2023 IEEE 20th international symposium on biomedical imaging (ISBI)*, pages 1–5. IEEE, 2023. 1

[26] Runqi Tang and Chongyang Ning. Mlfeu-net: A multi-scale low-level feature enhancement unet for breast lesions segmentation in ultrasound images. *Biomedical Signal Processing and Control*, 100:106931, 2025. 2

[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. 2

[28] Xiaoyang Wang, Bingfeng Zhang, Limin Yu, and Jimin Xiao. Hunting sparsity: Density-guided contrastive learning for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3114–3123, 2023. 2

[29] Yuchao Wang, Haochen Wang, Yujun Shen, Jingjing Fei, Wei Li, Guoqiang Jin, Liwei Wu, Rui Zhao, and Xinyi Le. Semi-supervised semantic segmentation using unreliable pseudo-labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4248–4257, 2022. 2, 6

[30] Linshan Wu, Leyuan Fang, Xingxin He, Min He, Jiayi Ma, and Zhun Zhong. Querying labeled for unlabeled: Cross-image semantic consistency guided semi-supervised semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):8827–8844, 2023. 2, 6

[31] Ruichao Wu, Xiangyu Lu, Zihuan Yao, and Yide Ma. Mfm-snet: A multi-frequency and multi-scale interactive cnn-transformer hybrid network for breast ultrasound image segmentation. *Computers in Biology and Medicine*, 177: 108616, 2024. 2, 6

[32] Min Xian, Yingtao Zhang, Heng-Da Cheng, Fei Xu, Boyu Zhang, and Jianrui Ding. Automatic breast ultrasound image segmentation: A survey. *Pattern Recognition*, 79:340–355, 2018. 1

[33] Xiaozheng Xie, Jianwei Niu, Xuefeng Liu, Yong Wang, Qingfeng Li, and Shaojie Tang. A domain knowledge powered hybrid regularization strategy for semi-supervised breast cancer diagnosis. *Expert Systems with Applications*, 243:122897, 2024. 2, 6

[34] Cheng Xue, Lei Zhu, Huazhu Fu, Xiaowei Hu, Xiaomeng Li, Hai Zhang, and Pheng-Ann Heng. Global guidance network for breast lesion segmentation in ultrasound images. *Medical image analysis*, 70:101989, 2021. 1

[35] 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. 2

[36] Lihe Yang, Lei Qi, Litong Feng, Wayne Zhang, and Yinghuan Shi. Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7236–7246, 2023. 1, 2, 6

[37] Ziyun Yang and Sina Farsiu. Directional connectivity-based segmentation of medical images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11525–11535, 2023. 3

[38] Ziyun Yang, Somayyeh Soltanian-Zadeh, and Sina Farsiu. Biconnet: An edge-preserved connectivity-based approach for salient object detection. *Pattern recognition*, 121:108231, 2022. 3

[39] Moi Hoon Yap, Manu Goyal, Fatima Osman, Robert Martí, Erika Denton, Arne Juette, and Reyer Zwiggelaar. Breast ultrasound region of interest detection and lesion localisation. *Artificial Intelligence in Medicine*, 107:101880, 2020. 2, 6

[40] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in Neural Information Processing Systems*, 34:18408–18419, 2021. 2

[41] Zhen Zhao, Sifan Long, Jimin Pi, Jingdong Wang, and Luping Zhou. Instance-specific and model-adaptive supervision for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23705–23714, 2023. 1

[42] Zhen Zhao, Lihe Yang, Sifan Long, Jimin Pi, Luping Zhou, and Jingdong Wang. Augmentation matters: A simple-yet-effective approach to semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11350–11359, 2023. 2, 6

[43] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, pages 3–11. Springer, 2018. 2

[44] 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. 2