IntraSense-Net: A Breast Tumor Ultrasound Image Segmentation Method with Semantic Differentiation

Jiayan Dai Zhihui Lai
Shenzhen University
Shenzhen, China
2310273125@email.szu.edu.cn lai_zhi_hui@163.com

Heng Kong
Shenzhen Nanshan Hospital
Shenzhen, China
generaldoc@126.com

Abstract

In medical ultrasound images, the phenomenon of "identical texture but distinct semantics" is frequently observed, where certain regions exhibit highly similar textures but correspond to significantly different semantic labels. This poses a critical challenge for traditional segmentation methods based on local convolutional features, as such methods are easily misled by texture similarity.To address this issue, we propose a problem-oriented segmentation framework, named IntraSense-Net, designed to alleviate semantic ambiguity under high texture similarity in breast ultrasound images.First, a global–local collaborative perception mechanism establishes long-range dependencies via non-local operations while leveraging spatial attention to fuse global statistical representations with local detailed responses, thereby constructing cross-region semantic associations.In addition, an entropy-based dynamic feature normalization strategy is introduced to mitigate error propagation in uncertain regions.Furthermore, a channel decoupling and re-weighting mechanism is employed to suppress texture-induced channel redundancy through orthogonal constraints, enabling more discriminative feature responses.The proposed IntraSense-Net achieves state-of-the-art performance across all key metrics on three public datasets, namely DatasetB, BUSI, and DDTI.

Keywords: ***ultrasound segmentation,global-local collaborative perception,entropy-based normalization,channel decoupling; attention mechanism***

1. Introduction

Ultrasound imaging has become an effective modality for early breast cancer screening due to its non-invasive nature, absence of ionizing radiation, and cost-effectiveness. Accurate segmentation of tumor regions in breast ultrasound (BUS) images plays a crucial role in clinical diagnosis, as it enables physicians to better assess lesion characteristics and disease progression. However, the challenges of BUS image segmentation arise not only from image complexity, but also from the intricate interplay between the physical ultrasound imaging mechanism and the structural properties of breast tissue. This interplay gives rise to the phenomenon commonly referred to as "identical texture but distinct semantics."

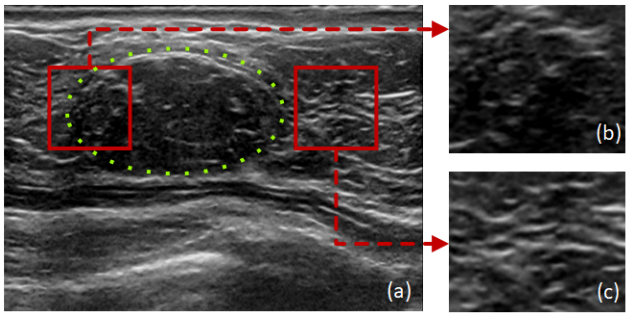

Figure 1. (a)shows the high similarity in texture between the tumor region (b) and the background region (c) in breast ultrasound images. Although they are nearly indistinguishable in terms of texture, their semantic information differs significantly, posing a major challenge in distinguishing the tumor region from the background. The red box highlights areas of texture similarity between the tumor and background regions, while the green dashed line marks the location of the tumor.

During ultrasound imaging, reflections at tissue interfaces frequently generate artifacts, causing tumor regions to exhibit texture patterns that are highly similar to surrounding normal tissue. More critically, intrinsic biological characteristics of tumors, such as the heterogeneous distribution of microcalcifications and fibrotic components, further contribute to echo patterns within lesions that closely resemble background tissue. As illustrated in Fig. 1, certain tumor regions in BUS images share nearly indistinguishable local textures with background areas, despite representing fundamentally different semantic categories. This phenomenon poses a substantial challenge for existing deep learning models that rely on pixel-wise predictions.

When local texture features dominate the representation,

shallow convolutional features are prone to being misled by texture similarity, resulting in incorrect classification. Conversely, deeper feature representations, although capable of encoding more abstract semantic information, often suffer from the loss of fine-grained discriminative cues that are essential for precise boundary delineation. The coexistence of strong local texture similarity and inconsistent global semantic cues makes it difficult for segmentation models to jointly preserve discriminative details and maintain reliable semantic separation, ultimately degrading segmentation performance.

In recent years, convolutional neural network (CNN)-based methods have achieved remarkable progress in medical image segmentation. For example, U-Net++ [43] enhances detail preservation through nested skip connections, while DeepLabv [6] expands the receptive field using atrous convolutions. Nevertheless, these approaches still exhibit inherent limitations when addressing the "identical texture but distinct semantics" problem. First, due to the locality of convolution operations, models often struggle to establish semantic associations across spatially distant regions, leading to confusion among areas with similar texture patterns. Second, the progressive loss of spatial resolution in deep feature representations compromises the accurate localization of small or irregular tumor regions.

Although hybrid architectures that integrate Transformers with CNNs have demonstrated improved global context modeling capability, several studies have reported that such architectures may exhibit sensitivity to noise and intensity variations in ultrasound imaging scenarios, which can adversely affect robustness under severe speckle interference. These observations highlight a fundamental issue in BUS image segmentation: during feature learning, texture similarity and semantic discrepancy are often tightly intertwined, while existing methods lack explicit mechanisms to decouple their influence. As a result, systematic misclassification may occur in regions exhibiting identical textures but distinct semantic meanings.

To address this challenge, we propose a problem-oriented dual-path collaborative perception framework to alleviate semantic ambiguity caused by highly similar textures in breast ultrasound images. First, a global–local collaborative perception strategy is adopted, where a non-local operator establishes long-range dependencies to capture the overall distribution pattern of tumors, while spatial attention adaptively integrates global statistical information with local detail responses, thereby facilitating cross-region semantic disambiguation. In addition, an entropy-based dynamic feature normalization strategy is incorporated to differentiate high- and low-confidence regions, aiming to stabilize feature propagation and mitigate error amplification. Second, a channel decoupling and re-weighting strategy is introduced to suppress channel correlations induced by texture similarity through orthogonal constraints, enabling more discriminative feature responses. By working in synergy, these two components collaboratively disentangle the coupling between texture and semantics, optimize feature representation, and improve segmentation robustness.

Our main contributions are as follows:

- We propose a problem-oriented global–local collaborative perception strategy tailored to semantic ambiguity in breast ultrasound images, which integrates long-range dependency modeling with local detail enhancement to support cross-region semantic disambiguation.
- We introduce a channel decoupling and re-weighting strategy to suppress texture-induced channel redundancy and promote discriminative feature representation through collaborative channel optimization.
- We conduct systematic experiments on three public datasets, including BUS, BUSI, and DDTI. The experimental results demonstrate that the proposed framework consistently outperforms traditional U-Net and other state-of-the-art methods across multiple evaluation metrics, validating its effectiveness and stability.

2. Related works

2.1. Medical image segmentation architecture

Medical image segmentation is a key task in the medical field. U-Net [29] introduced an encoder–decoder structure with skip connections to mitigate detail loss, enabling end-to-end training with a small number of annotated samples. V-Net [24] further extended the U-Net structure to three-dimensional medical images and employed a Dice coefficient-based loss function to address the foreground–background voxel imbalance issue. However, U-Net's single-path feature fusion approach has limitations when dealing with complex texture and semantic conflicts in ultrasound images. Therefore, UNet++ [43], CE-Net [11], MSU-Net [31], and EMCAD [28] have introduced techniques such as multi-scale feature fusion, dense connections, residual modules, and attention mechanisms, achieving excellent performance in various medical image segmentation tasks.

The Transformer hybrid architecture provides new insights for medical image segmentation. TransUNet [5] integrates the Transformer into the U-Net encoder and decoder, surpassing nn-UNet [16] in multi-organ and pancreatic tumor segmentation tasks. Swin-Unet [2] enhances contextual features through a window-based self-attention mechanism, while UCTNet [12] optimizes the complementarity of Transformer and CNN through uncertainty estimation. CSWin-UNet [20] improves computational efficiency using the CSWin self-attention mechanism. HAU-Net [40] integrates long-range dependencies and local detail modeling

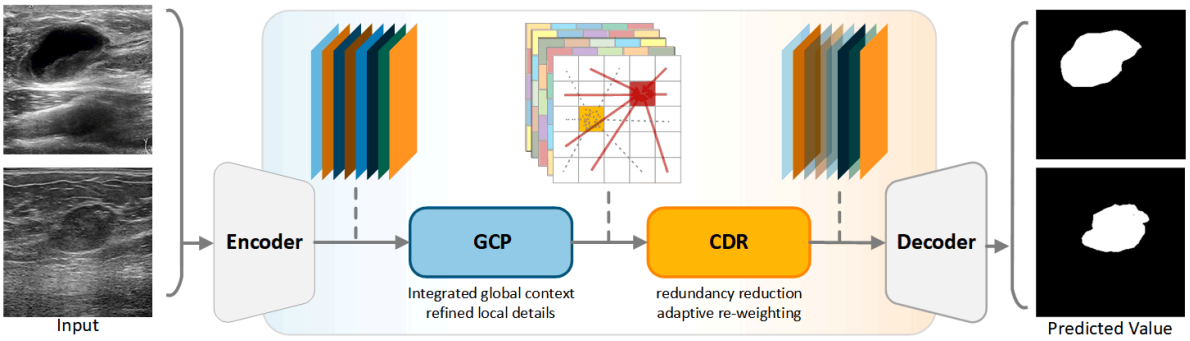

Figure 2. The diagram of the overall architecture of IntraSense-Net. The model employs an encoder to extract initial features, followed by the global-local collaborative perception module (GCP) to integrate global and local information to enhance feature representation capability. Then, the channel decoupling and re-weighting module (CDR) optimizes feature distribution to improve channel discriminability. Finally, the decoder progressively restores spatial resolution to generate high-precision segmentation results.

for breast ultrasound images, demonstrating superior performance across multiple public datasets. Despite these advances, several studies have reported that Transformer-based or hybrid architectures may exhibit sensitivity to noise and intensity variations in ultrasound imaging scenarios, which can affect their robustness under heavy speckle interference.

In response to the segmentation requirements of ultrasound images, researchers have proposed different solutions. For instance, [22] incorporates attention mechanisms to enhance breast tumor classification performance, while MEF-UNet [38] designs a selective feature extraction module to address the complexity of lesions. The BaS model [26] achieves efficient breast lesion segmentation in resource-constrained environments, and NU-Net [4] further optimizes breast ultrasound segmentation by sharing U-Net weights at different depths.

2.2. Improvement on attention mechanism

In medical image segmentation tasks, attention mechanisms play a crucial role in enhancing a model's ability to recognize complex structures, especially in ultrasound image segmentation. Squeeze-and-Excitation Networks [13] improve feature representation through channel recalibration, CBAM [36] combines channel and spatial attention for adaptive feature refinement, and the Dual Attention Network [8] further models semantic relationships using both position and channel attention. CCNet [15] efficiently captures global dependencies via a criss-cross attention mechanism and achieves excellent performance across various visual tasks.

To improve segmentation performance, researchers have proposed a series of specialized attention mechanisms. Medical Transformer [33] combines gated axial attention with a local–global training strategy, and Prior Attention Network [42] enhances multi-lesion segmentation via lesion-related spatial attention. Channel Prior Convolutional Attention [14] dynamically allocates attention weights across both channel and spatial dimensions. CA-Net [10] integrates multiple attention mechanisms to improve segmentation accuracy and interpretability, while TransAttUnet [3] optimizes multi-scale context feature learning through multi-level guided attention. These methods have demonstrated superior performance compared to existing approaches in medical image segmentation tasks. However, these attention-based methods are generally designed to enhance feature representation, rather than explicitly disentangle semantic ambiguity arising from highly similar textures in ultrasound images.

2.3. Feature decoupling technique

Feature disentanglement techniques aim to separate different feature types, which has profound implications for medical image analysis [9, 23]. FactorVAE [18] enhances disentanglement through factorized representations. The Adaptive Disentangled Transformer [41] optimizes cross-layer disentanglement by constraining the independence of attention heads. DecoupleSegNet [37] combines body and edge features to improve the performance of osteosarcoma MRI segmentation. In terms of channel decoupling, references [7, 19, 44, 35, 30] employ variational autoencoders and global–local feature decoupling methods to remove source-specific information, thereby enhancing unsupervised domain adaptation capabilities. These studies suggest that decoupling learning plays a crucial role in enhancing the generalization ability of medical image segmentation. Nevertheless, most existing decoupling techniques are not specifically designed to suppress texture-induced channel redundancy for semantic disambiguation in breast ultrasound image segmentation.

3. Methodology

In this study, we present IntraSense-Net, a breast tumor ultrasound image segmentation framework designed to ad-

dress semantic ambiguity in ultrasound imaging. The proposed method targets the "identical texture but distinct semantics" problem from two complementary perspectives. On the one hand, by jointly modeling global contextual information and local structural details, the network is encouraged to capture long-range semantic dependencies while preserving fine-grained discriminative cues. On the other hand, by explicitly maintaining diversity among channel representations, the model suppresses feature homogenization caused by texture redundancy, enabling different channels to encode complementary semantic information.

Based on these considerations, we construct a dual-path collaborative perception framework in which global–local perception and channel decoupling are alternately integrated within the feature extraction process. Each component is designed to address a specific aspect of semantic confusion, and their coordinated interaction facilitates more stable and discriminative feature learning, ultimately improving segmentation robustness and accuracy.

Specifically, as illustrated in Fig. 2, a ResNet50 backbone is first employed as the encoder to extract initial feature representations, which capture rich local texture patterns and coarse semantic information. Subsequently, a global–local collaborative perception module is applied to integrate long-range contextual cues with local detail responses, thereby enhancing cross-region semantic consistency. Following this, a channel decoupling and reweighting module is introduced to suppress texture-induced channel correlations through orthogonal constraints, promoting discriminative and diverse channel-wise representations. Finally, the decoder progressively restores spatial resolution and refines the fused features to generate precise pixel-level segmentation results.

3.1. Global–local collaborative perception mechanism

To alleviate semantic ambiguity caused by highly similar textures in breast ultrasound images, we employ a global–local collaborative perception mechanism (GCP), as illustrated in Fig. 3. The core motivation of this design is to jointly model long-range contextual relationships and local structural details, enabling the network to establish cross-region semantic consistency while preserving fine-grained discriminative information.

Specifically, a non-local operator is first utilized to capture long-range dependencies, providing a global view of the tumor distribution pattern. Subsequently, a spatial attention mechanism is applied to adaptively integrate global statistical information with local detail responses, facilitating the propagation of semantic cues across spatially distant regions. In addition, an entropy-based dynamic feature normalization strategy is incorporated to differentiate high-confidence and low-confidence feature regions, thereby stabilizing feature propagation and mitigating error amplification.

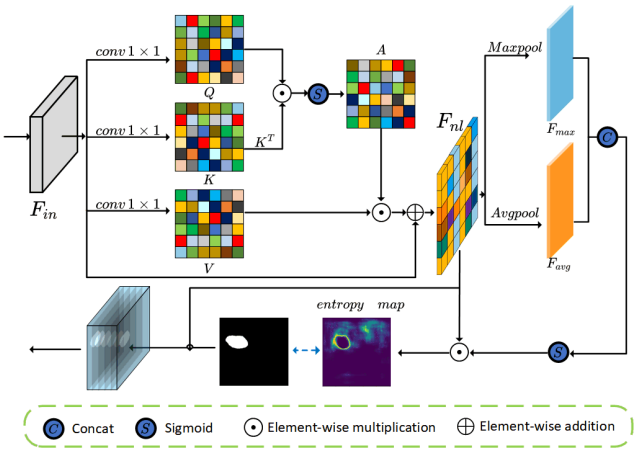

Figure 3. Global-local collaborative perception module. This module integrates global information and local details to help the model more accurately locate tumor regions, while also enhancing its ability to recognize complex backgrounds.

tion. Through this collaborative design, the GCP module enhances the discriminative capability of feature representations under challenging texture-similarity conditions.

First, for the input feature map $F_{\text{in}} \in \mathbb{R}^{C\times H\times W}$, we apply a 1×1 convolution to map it to different representation spaces, generating queries $\mathbf{Q}$, keys $\mathbf{K}$, and values $\mathbf{V}$, as follows 1:

$$\mathbf{Q} = W_{\mathbf{Q}}F_{\text{in}}, \quad \mathbf{K} = W_{\mathbf{K}}F_{\text{in}}, \quad \mathbf{V} = W_{\mathbf{V}}F_{\text{in}} \tag{1}$$

where $W_{\mathbf{Q}}, W_{\mathbf{K}}, W_{\mathbf{V}} \in \mathbb{R}^{C\times C'}$ are learnable weight matrices, and C' denotes the dimension after the mapping. This provides a more compact feature representation for subsequent global information modeling.

Next, we adopt a global dependency modeling approach to compute the similarity between queries and keys and construct a global affinity matrix as formulated in 2.

$$\mathbf{A} = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right) \tag{2}$$

where d is the dimension of the key vector. The Softmax normalization ensures that the sum of relationship weights at each position equals 1. The matrix $\mathbf{A} \in \mathbb{R}^{(HW)\times(HW)}$ models the relationships between all spatial positions in the image, enabling each position not only rely on local features but also benefit from global information to gain a more comprehensive semantic understanding. Thus, global information is incorporated into the original features through weighted summation, and initial details are retained via residual connections, i.e 3:

$$\mathbf{F}_{\text{nl}} = \mathbf{A}\mathbf{V} + \mathbf{F}_{\text{in}} \tag{3}$$

After the global information interaction, the features of each position are supplemented with information from the

global scope, thus overcoming the limitations of traditional convolution, restricted to fixed local windows.

However, directly performing global fusion may introduce noise, weakening the local discriminative information. To address this, we further apply a spatial attention mechanism on the globally fused features F_{nl}. First, we perform global average pooling and max pooling on F_{nl} to obtain two complementary global statistics, F_{avg} and F_{max}. After concatenating, we pass them through a convolutional layer followed by a Sigmoid activation to generate the spatial attention map 4.

$$A = \sigma\left(\text{Conv}\left([F_{\text{avg}}, F_{\text{max}}]\right)\right) \tag{4}$$

The spatial attention map reflects the importance of each spatial position in terms of its discriminative power, thereby selectively enhancing relevant information. Next, we element-wise multiply the attention map with the globally fused features F_{nl} 5.

$$F_{\text{att}} = A_{\text{spatial}} \odot F_{\text{nl}} \tag{5}$$

This operation ensures that the critical local regions for the segmentation task are significantly enhanced, while irrelevant or noisy information is effectively suppressed.

To ensure the stability of features during the information transmission process, we design an entropy-based region whitening strategy. Specifically, we compute the entropy value of the normalized activation probabilities of each channel at each spatial position (x, y) as follow 6:

$$H(x, y) = -\sum_{c=1}^{C} p_c(x, y) \log p_c(x, y) \tag{6}$$

Low entropy regions usually indicate concentrated and more certain feature distributions. Therefore, we construct a binary mask $M(x, y) = (H(x, y) < \tau)$ with a preset threshold $\tau = 0.5$ and apply normalization to these low-entropy regions 7.

$$\begin{aligned} F_{\text{white}}(x, y) = M(x, y) \cdot \left(\frac{F_{\text{att}}(x, y) - \mu}{\sigma}\right) \\ + \left(1 - M(x, y)\right) \cdot F_{\text{att}}(x, y) \end{aligned} \tag{7}$$

where μ and σ represent the local mean and standard deviation.

3.2. Channel decoupling and re-weighting

To further enhance the discriminative quality of feature representations, we incorporate a channel decoupling and re-weighting mechanism (CDR), as illustrated in Fig. 4. The motivation of this design is to mitigate representation homogenization caused by texture similarity, which often induces high redundancy among channel activations in breast ultrasound images.

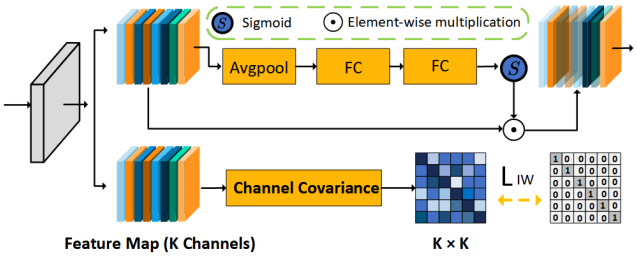

Figure 4. Channel decoupling and re-weighting module. This module optimizes feature representation by decoupling the correlations between channels and re-weighting the features, thereby enhancing the model's ability to focus on tumor regions and reducing the interference from redundant information.

Specifically, a channel decoupling strategy is first applied to reduce inter-channel correlations, encouraging different channels to encode complementary information. Subsequently, a channel attention mechanism is employed to adaptively re-weight each channel based on its global response, thereby restoring the unique contribution of individual channels to the overall feature representation. Through this process, the CDR module suppresses texture-induced redundancy and promotes diverse and discriminative channel-wise representations, facilitating more reliable semantic differentiation.

Let the input feature be $F \in \mathbb{R}^{C \times H \times W}$. To alleviate the activation redundancy caused by similar textures, we introduce instance whitening, which transforms the feature F such that each channel can capture more independent information. The transformation is defined as equation 8:

$$F_{\text{decouple}} = F - \text{Proj}_{\text{span}(F)}(F) \tag{8}$$

where $\text{Proj}_{\text{span}(F)}(F)$ projects F onto its own subspace. By increasing the rank of the feature matrix, the redundant responses are reduced.

To optimize instance whitening, we introduce a loss function L_{IW} that preserves useful information while suppressing redundancy, using the following 9.

$$\begin{aligned} L_{\text{IW}} = \alpha \cdot \left(\text{rank}(F) - \text{rank}(F_{\text{decouple}})\right) \\ + \beta \cdot \text{dist}(F, F_{\text{decouple}}) \end{aligned} \tag{9}$$

where $\text{rank}(F)$ and $\text{rank}(F_{\text{decouple}})$ represent the ranks of the features, $\text{dist}(F, F_{\text{decouple}})$ is the distance measure between them, and α and β are weight coefficients. By minimizing this loss function, redundancy is reduced, and the model's feature representation ability is enhanced.

Based on this, we further re-weight the decoupled features by channel. First, we extract statistical descriptions for each channel through global average pooling (see Equation 10).

$$s_c = \frac{1}{H \times W} \sum_{x=1}^{H} \sum_{y=1}^{W} F_c(x, y), \quad c = 1, 2, \ldots, C \tag{10}$$

This statistical vector $s = [s_1, s_2, ..., s_C]$ reflects the contribution of each channel in the global feature space. Next, we input s into a multilayer perceptron (MLP), which performs a nonlinear mapping to generate normalized channel attention weights (see Equation 11).

$$w_c = \sigma(\text{MLP}(s_c)) \tag{11}$$

where $\sigma(\cdot)$ is the Sigmoid activation function, ensuring the weights are constrained within the range of [0, 1]. Finally, we use these weights to re-weight the decoupled features channel-wise, yielding the final output (see Equation 12).

$$F_{\text{out},c} = w_c \cdot F_{\text{decouple},c} \tag{12}$$

Through the decoupling operation, redundancy caused by texture similarity is suppressed, resulting in a feature space with higher rank. Then, through adaptive reweighting, the importance of each channel is reassessed, highlighting the features that are discriminative for the segmentation task.

3.3. Loss function

In this study, the loss function consists of three components: Cross-Entropy Loss, Dice Loss, and Instance Whitening Loss.The final total loss function is the weighted sum of these three losses (see Equation 13).

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{Dice}} + \mathcal{L}_{\text{IW}} \tag{13}$$

where $\mathcal{L}_{\text{CE}}$ and $\mathcal{L}_{\text{Dice}}$ are defined as follows (see Equation 14, 15).

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N} \sum_{i=1}^{N} y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \tag{14}$$

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2 \sum_{i=1}^{N} y_i \hat{y}_i}{\sum_{i=1}^{N} y_i + \sum_{i=1}^{N} \hat{y}_i} \tag{15}$$

where y_i is the true label, $\hat{y}_i$ is the predicted probability, and N is the total number of samples.

4. Experiments and results

4.1. Implementation details

In this study, experiments were conducted using the PyTorch framework on an NVIDIA RTX 3090 GPU. To comprehensively validate the effectiveness of the proposed method, experiments were performed on three public datasets. To ensure consistent data distribution, both benign and malignant samples were evenly distributed across the datasets. During the experiment, all images were resized to 256×256 pixels. Data augmentation techniques, including image flipping and rotation, were applied to prevent overfitting. For training, a batch size of 4 was used, and the RMSprop optimizer with an initial learning rate of 1×10^{-4} was adopted. Additionally, weight decay was set to 1×10^{-8}. To evaluate the proposed method, 5-fold cross-validation was employed across all approaches.

4.2. Evaluation metrics

To thoroughly and objectively evaluate the segmentation performance of the proposed method, we select evaluation metrics that are highly relevant to different segmentation tasks, namely Dice Similarity Coefficient (Dice), Jaccard Coefficient (Jaccard), Accuracy, Recall, and Precision. These metrics allow for a comprehensive assessment of the method's effectiveness in various aspects of segmentation quality.

4.3. Dataset

This study evaluates the performance of the segmentation network using three widely used public datasets.

The first dataset is the Dataset B [1] collected by Yap et al., which contains 163 breast ultrasound images. The average resolution of the images is 760×570 pixels, and the dataset includes two categories: benign lesions and cancerous masses.

The second dataset is the BUSI dataset [39], which contains 780 breast ultrasound images with an average resolution of 500×500 pixels. The images are classified into three categories: normal, benign, and malignant. In this study, normal cases were not used for training and testing the network, as the main goal of the research is to segment lesions.

The third dataset is the DDTI dataset [27], provided by the National University of Colombia, Cim@Lab, and IDIME (Institute for Medical Diagnosis). It consists of 683 thyroid nodule ultrasound images with an average resolution of 500×500 pixels. The dataset is divided into benign and malignant categories.

4.4. Contrast experiment

In this study, we conducted a comprehensive evaluation of the proposed method on three public datasets: Dataset B, BUSI, and DDTI. The comparison methods were selected based on their public availability, reproducibility, and established effectiveness in breast ultrasound image segmentation tasks, covering both classical CNN-based models and recent advanced architectures.The proposed method was compared with advanced methods such as U-Net [29], EA-Net [34], CMUNet [32], MicroSegNet [17], RollingUnet [21], EBTNet [25], and EMCAD [28]. The

Model	Dice (%)	Jaccard (%)	Accuracy (%)	Recall (%)	Precision (%)
U-Net	66.76	53.79	96.95	84.51	63.68
EA-Net	77.92	67.68	97.60	84.17	78.99
CMUNet	79.09	70.29	97.77	79.06	83.13
MicroSegNet	<u>83.38</u>	<u>75.38</u>	<u>98.28</u>	84.48	<u>87.22</u>
RollingUnet	82.58	72.66	98.06	<u>88.60</u>	80.45
EBTNet	79.59	70.19	98.00	80.30	82.57
EMCAD	80.04	71.00	98.20	83.43	82.14
Our	**87.13**	**78.94**	**98.40**	**88.68**	**88.97**

Table 1. Performance comparison of different methods on the Dataset B. Bold indicates the best result for the corresponding metric, and underscore indicates the second-best result.

Model	Dice (%)	Jaccard (%)	Accuracy (%)	Recall (%)	Precision (%)
U-Net	66.72	55.35	95.08	79.15	67.79
EA-Net	77.56	66.33	94.08	**85.23**	77.27
CMUNet	75.27	63.72	93.72	80.88	76.87
MicroSegNet	79.43	<u>68.95</u>	<u>94.53</u>	82.43	81.48
RollingUnet	78.84	67.65	93.96	85.03	78.84
EBTNet	76.54	65.37	94.06	81.83	78.14
EMCAD	<u>80.12</u>	68.02	94.01	<u>85.19</u>	<u>82.66</u>
Our	**81.60**	**71.31**	**95.45**	81.57	**85.30**

Table 2. Performance comparison of different methods on the BUSI. Bold indicates the best result for the corresponding metric, and underscore indicates the second-best result.

experimental results consistently demonstrated competitive or superior performance across multiple evaluation metrics, validating the effectiveness of the global-local collaborative perception and channel decoupling re-weighting strategies in addressing the "same texture, different semantics" issue.

As shown in the results from Table 1, Table 2, and Table 3, our method demonstrates clear advantages across multiple datasets. Compared to U-Net, on Dataset B, the Dice coefficient improves from 66.76% to 87.13%, and on the BUSI dataset, it increases from 52.19% to 81.60%, validating its robustness on complex ultrasound images. Moreover, our method achieves a better balance between Recall and Precision, reducing both false negatives and false positives, and ensuring higher-quality segmentation results.

In comparison to other advanced methods, our model also performs competitively. For instance, on Dataset B, MicroSegNet achieves a Precision of 87.22%, while our method reaches 88.97%, indicating stronger false-positive suppression, which is particularly important for medical image segmentation. On the BUSI dataset, although RollingUnet achieves a Recall of 85.03%, its Precision is 78.84%, whereas our method achieves 85.30%, enabling more precise segmentation and reducing both false positives and false negatives.

On the DDTI dataset, our method achieves a Dice score of 81.41% and a Jaccard index of 72.91%, and its Precision reaches 83.71%, which is slightly higher than EMCAD's 83.38%, further demonstrating its capability in handling low-contrast and ambiguous regions.

In summary, the experimental results across all three datasets clearly demonstrate that the method proposed in this paper significantly outperforms traditional U-Net and other state-of-the-art methods in key metrics such as Dice, Jaccard, Accuracy, Recall, and Precision. Overall, the proposed combined strategy not only provides a systematic and effective solution to the "same texture, different semantics" problem in theory but also exhibits outstanding segmentation performance in practice.

4.5. Ablation experiment

To investigate the contribution of each component of the model, we designed a series of ablation experiments to analyze the impact of different modules on segmentation performance. The experimental results are shown in 4.

The baseline achieved a Dice rate of 79.74% and a Jaccard index of 69.94% on the BUS dataset. Although it is able to extract some representative features, it still faces challenges in segmenting complex ultrasound images. In particular, the Recall and Precision values are relatively modest, indicating that the model struggles to accurately distinguish lesion regions from background. Furthermore, when facing significant intra-class variations, the feature representation is prone to interference, leading to mis-segmentation. This suggests that relying solely on the

Model	Dice (%)	Jaccard (%)	Accuracy (%)	Recall (%)	Precision (%)
U-Net	67.12	53.75	90.48	79.39	67.83
EA-Net	74.78	66.30	96.02	76.72	79.07
CMUNet	73.93	64.47	96.20	73.39	80.16
MicroSegNet	78.15	69.39	96.27	78.13	82.68
RollingUnet	76.39	66.34	95.72	76.94	80.26
EBTNet	74.33	64.40	96.13	74.70	80.40
EMCAD	<u>79.99</u>	<u>70.81</u>	<u>96.60</u>	<u>80.82</u>	<u>83.38</u>
Our	**81.41**	**72.91**	**97.03**	**82.14**	**83.71**

Table 3. Performance comparison of different methods on the DDTI dataset. Bold indicates the best result for each metric, and underline indicates the second-best result.

GCP	CDR	Dice (%)	Jaccard (%)	Accuracy (%)	Recall (%)	Precision (%)
		79.74	69.94	97.97	83.16	83.94
✔†		84.43	75.64	98.20	86.81	87.12
✔		85.10	76.01	98.21	86.76	87.46
	✔	85.33	76.26	98.20	88.10	86.53
✔	✔	87.13	78.94	98.40	88.68	88.97

Table 4. Performance evaluation of different modules in the ablation study. † indicates a simplified GCP variant without entropy-based dynamic feature normalization (GCP–EN).

multi-scale feature extraction of the backbone network still has limitations in terms of information discrimination.

After introducing the global-local collaborative perception module, the Dice and Jaccard indices increased to 85.10% and 76.01%, respectively, with noticeable improvements in Accuracy, Recall, and Precision. This demonstrates that the module, with the aid of global context information, enables more effective parsing of complex tissue structures, allowing the model to more accurately capture the feature distribution of lesion regions. Additionally, it helps reduce misclassification caused by the similarity of local features, ensuring that the model maintains high discriminative ability even in complex backgrounds.We further evaluate a simplified GCP variant without entropy-based dynamic feature normalization, which results in a measurable performance drop, indicating that uncertainty-aware normalization plays a critical role in stabilizing feature propagation under texture ambiguity.

On the other hand, only when the channel decoupling and re-weighting module is introduced, the Dice and Jaccard indices reach 85.33% and 76.26%, respectively. Compared to the global-local collaborative perception module, it performs better in Recall (88.10%) but shows a slight decrease in Precision. This indicates that the module plays a positive role in enhancing channel information independence and reducing redundant features, allowing the model to focus more on the feature representation of lesion regions. However, without the guidance of global features, the model may be influenced by locally highly similar regions, which could lead to incorrect activations in certain cases.

When both modules are used together, the model achieves best performance across all metrics, with the Dice and Jaccard indices improving to 87.13% and 78.94%, respectively. Accuracy also increases to 98.40%, while Recall and Precision reach 88.68% and 88.97%. This result validates the synergistic effect of the two modules: the global-local collaborative perception module enhances the overall feature integrity, improving the model's ability to understand complex organizational structures, while the channel decoupling and re-weighting module optimizes feature representation, allowing the model to more effectively differentiate between lesion regions and background noise. The combination of both modules enables the model to not only accurately identify lesion areas but also maintain stable segmentation performance when facing intra-class variations, ultimately improving the model's generalization capability and robustness.

4.6. Visual analysis

To further validate the segmentation performance of different methods on breast ultrasound images, we selected various types of lesions for visual analysis. Figure 5 shows the segmentation results of different methods on typical cases, where it is evident that the performance of each method varies significantly when facing different challenges.

Firstly, for lesions with clear boundaries and uniform internal textures, all methods are able to perform the segmentation task relatively well, indicating that most segmenta-

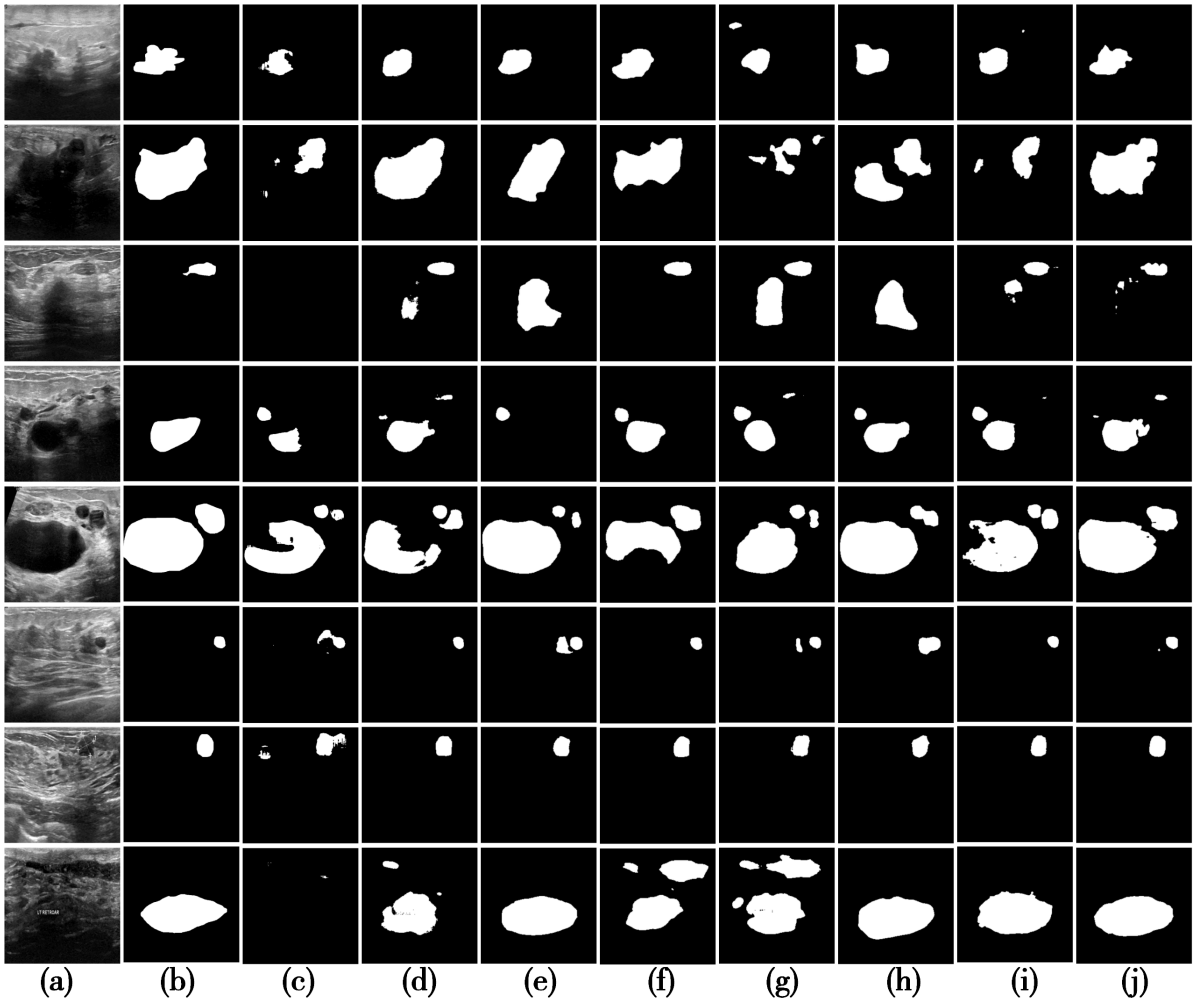

Figure 5. Visualization comparison of different methods in ultrasound image segmentation. (a) Original image, (b) Ground truth, (c)-(j) represent the segmentation results of (c) U-Net, (d) EA-Net, (e) CMUNet, (f) MicroSegNet, (g) RollingUnet, (h) EBTNet, (i) EMCAD, (j) our method.

tion networks can stably learn the morphological features of lesions in simpler cases. However, when the lesions exhibit internal texture variations, the results start to diverge. In this scenario, both EBTNet and our method still manage to accurately segment the lesion areas, while other methods exhibit shape distortion or boundary loss in some samples. This suggests that these methods have limited generalization ability when dealing with lesions with complex internal structures.

Secondly, for lesions with texture features similar to the background and prone to confusion, our method demonstrates the strongest robustness. Regardless of the sample, its segmentation results remain consistently stable and accurate. In contrast, other methods show instability, performing well on some samples while suffering from significant missegmentation or missed detection on others. This indicates that the method we propose has a stronger discriminative ability when addressing the ”same texture but different semantics” problem.

Finally, for the segmentation of malignant tumors, all methods are generally able to locate the approximate position of the lesions. However, our method stands out in terms of shape preservation and boundary detail delineation. This suggests that our approach not only captures large-scale features of the lesions effectively but also excels in expressing finer details.

The experimental results demonstrate that our method performs consistently well across various challenging scenarios, enabling more accurate extraction of lesion regions. In particular, it shows superior performance in identifying confusing lesions and maintaining their shape better when compared to other methods.

5. Conclusion

The method proposed in this study effectively addresses the "same texture but different semantics" issue in ultrasound images by associating long-range dependency information with local details and decoupling the correlation between channels. This enhances the accuracy of tumor region segmentation. Compared to traditional methods, the model demonstrates significant improve method in feature distribution optimization and information fusion, especially in stable segmentation of different types of tumors. The visualization results show that the model maintains high-precision segmentation even in high-similarity backgrounds and complex texture interference, proving its applicability in real clinical environments.

Although this study has made significant progress, there are several research directions worth further exploration. First, lightweight design and optimized feature extraction will help improve inference speed, making it more suitable for clinical applications. Second, cross-modal or multi-modal data fusion can enhance the robustness of the model, particularly when combining different imaging modalities, thus improving segmentation accuracy. In addition, weakly supervised or unsupervised learning methods can reduce the reliance on annotated data, thereby enhancing the model's generalization ability in scenarios with limited data.

The method proposed in this study demonstrates strong robustness and transferability, showcasing its potential in smart healthcare. With the advancement of artificial intelligence technologies, automated tumor detection and monitoring will become an integral part of intelligent medical systems. This method can effectively assist doctors in diagnostic work, improve efficiency, reduce workload, and provide strong support for precision medicine, thus promoting the application and development of smart healthcare.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China (Grant Nos. 62272319 and 62476175), in part by the Natural Science Foundation of Guangdong Province (Grant Nos. 2023A1515010677, 2024A1515011637, and 2023B1212060076), and in part by the Shenzhen Science and Technology Program (Grant Nos. JCYJ20220818095803007, JCYJ20240813142206009, and JCYJ20250604145234045).

References

[1] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy. Dataset of breast ultrasound images. *Data in brief*, 28:104863, 2020. 6

[2] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pages 205–218. Springer, 2022. 2

[3] B. Chen, Y. Liu, Z. Zhang, G. Lu, and A. W. K. Kong. Transattunet: Multi-level attention-guided u-net with transformer for medical image segmentation. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 8(1):55–68, 2023. 3

[4] G. Chen, L. Li, J. Zhang, and Y. Dai. Rethinking the unpretentious u-net for medical ultrasound image segmentation. *Pattern Recognition*, 142:109728, 2023. 3

[5] J. Chen, J. Mei, X. Li, Y. Lu, Q. Yu, Q. Wei, X. Luo, Y. Xie, E. Adeli, Y. Wang, et al. Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers. *Medical Image Analysis*, 97:103280, 2024. 2

[6] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 2

[7] W. Deng, L. Zhao, Q. Liao, D. Guo, G. Kuang, D. Hu, M. Pietikäinen, and L. Liu. Informative feature disentanglement for unsupervised domain adaptation. *IEEE Transactions on Multimedia*, 24:2407–2421, 2021. 3

[8] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3146–3154, 2019. 3

[9] Y. Gao, W. Xia, D. Hu, W. Wang, and X. Gao. Desam: Decoupled segment anything model for generalizable medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 509–519. Springer, 2024. 3

[10] R. Gu, G. Wang, T. Song, R. Huang, M. Aertsen, J. Deprest, S. Ourselin, T. Vercauteren, and S. Zhang. Ca-net: Comprehensive attention convolutional neural networks for explainable medical image segmentation. *IEEE transactions on medical imaging*, 40(2):699–711, 2020. 3

[11] Z. Gu, J. Cheng, H. Fu, K. Zhou, H. Hao, Y. Zhao, T. Zhang, S. Gao, and J. Liu. Ce-net: Context encoder network for 2d medical image segmentation. *IEEE transactions on medical imaging*, 38(10):2281–2292, 2019. 2

[12] X. Guo, X. Lin, X. Yang, L. Yu, K.-T. Cheng, and Z. Yan. Uctnet: Uncertainty-guided cnn-transformer hybrid networks for medical image segmentation. *Pattern Recognition*, 152:110491, 2024. 2

[13] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 3

[14] H. Huang, Z. Chen, Y. Zou, M. Lu, C. Chen, Y. Song, H. Zhang, and F. Yan. Channel prior convolutional attention for medical image segmentation. *Computers in Biology and Medicine*, 178:108784, 2024. 3

[15] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 603–612, 2019. 3

[16] F. Isensee, P. F. Jäger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein. Automated design of deep learning meth-

ods for biomedical image segmentation. *arXiv preprint arXiv:1904.08128*, 2019. 2

[17] H. Jiang, M. Imran, P. Muralidharan, A. Patel, J. Pensa, M. Liang, T. Benidir, J. R. Grajo, J. P. Joseph, R. Terry, et al. Microsegnet: A deep learning approach for prostate segmentation on micro-ultrasound images. *Computerized Medical Imaging and Graphics*, 112:102326, 2024. 6

[18] H. Kim and A. Mnih. Disentangling by factorising. In *International conference on machine learning*, pages 2649–2658. PMLR, 2018. 3

[19] D. Liu, C. Zhang, Y. Song, H. Huang, C. Wang, M. Barnett, and W. Cai. Decompose to adapt: Cross-domain object detection via feature disentanglement. *IEEE Transactions on Multimedia*, 25:1333–1344, 2022. 3

[20] X. Liu, P. Gao, T. Yu, F. Wang, and R.-Y. Yuan. Cswin-unet: Transformer unet with cross-shaped windows for medical image segmentation. *Information Fusion*, 113:102634, 2025. 2

[21] Y. Liu, H. Zhu, M. Liu, H. Yu, Z. Chen, and J. Gao. Rolling-unet: Revitalizing mlp's ability to efficiently extract long-distance dependencies for medical image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 3819–3827, 2024. 6

[22] Y. Luo, Q. Huang, and X. Li. Segmentation information with attention integration for classification of breast tumor in ultrasound image. *Pattern Recognition*, 124:108427, 2022. 3

[23] Z. Luo, Z. Jia, Z. Yuan, and J. Peng. Hdc-net: Hierarchical decoupled convolution network for brain tumor segmentation. *IEEE Journal of Biomedical and Health Informatics*, 25(3):737–745, 2020. 3

[24] F. Milletari, N. Navab, and S.-A. Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016. 2

[25] Y. Pan, L. Niu, X. Yang, Q. Niu, and B. Chen. Ebtnet: Efficient bilateral token mixer network for fetal cardiac ultrasound image segmentation. *IEEE Access*, 2024. 6

[26] Y. Pang, Y. Li, T. Huang, J. Liang, Z. Ding, H. Chen, B. Zhao, Y. Hu, Z. Zhang, and Q. Wang. Efficient breast lesion segmentation from ultrasound videos across multiple source-limited platforms. *IEEE Journal of Biomedical and Health Informatics*, 2025. 3

[27] L. Pedraza, C. Vargas, F. Narváez, O. Durán, E. Muñoz, and E. Romero. An open access thyroid ultrasound image database. In *10th International symposium on medical information processing and analysis*, volume 9287, pages 188–193. SPIE, 2015. 6

[28] M. M. Rahman, M. Munir, and R. Marculescu. Emcad: Efficient multi-scale convolutional attention decoding for medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11769–11779, 2024. 2, 6

[29] O. Ronneberger, P. Fischer, and T. 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, 6

[30] B. Shi, W. Li, J. Huo, P. Zhu, L. Wang, and Y. Gao. Global-and local-aware feature augmentation with semantic orthogonality for few-shot image classification. *Pattern Recognition*, 142:109702, 2023. 3

[31] R. Su, D. Zhang, J. Liu, and C. Cheng. Msu-net: Multi-scale u-net for 2d medical image segmentation. *Frontiers in Genetics*, 12:639930, 2021. 2

[32] F. Tang, L. Wang, C. Ning, M. Xian, and J. Ding. Cmu-net: a strong convmixer-based medical ultrasound image segmentation network. In *2023 IEEE 20th international symposium on biomedical imaging (ISBI)*, pages 1–5. IEEE, 2023. 6

[33] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel. Medical transformer: Gated axial-attention for medical image segmentation. In *Medical image computing and computer assisted intervention–MICCAI 2021: 24th international conference, Strasbourg, France, September 27–October 1, 2021, proceedings, part I 24*, pages 36–46. Springer, 2021. 3

[34] K. Wang, X. Zhang, X. Zhang, Y. Lu, S. Huang, and D. Yang. Eanet: Iterative edge attention network for medical image segmentation. *Pattern Recognition*, 127:108636, 2022. 6

[35] T. Wang, Z. Zheng, Z. Zhu, Y. Sun, C. Yan, and Y. Yang. Learning cross-view geo-localization embeddings via dynamic weighted decorrelation regularization. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 3

[36] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 3

[37] J. Wu, Y. Guo, F. Gou, and Z. Dai. A medical assistant segmentation method for mri images of osteosarcoma based on decouplesegnet. *International Journal of Intelligent Systems*, 37(11):8436–8461, 2022. 3

[38] M. Xu, Q. Ma, H. Zhang, D. Kong, and T. Zeng. Mef-unet: An end-to-end ultrasound image segmentation algorithm based on multi-scale feature extraction and fusion. *Computerized Medical Imaging and Graphics*, 114:102370, 2024. 3

[39] M. H. Yap, M. Goyal, F. Osman, R. Martí, E. Denton, A. Juette, and R. Zwiggelaar. Breast ultrasound region of interest detection and lesion localisation. *Artificial intelligence in medicine*, 107:101880, 2020. 6

[40] H. Zhang, J. Lian, Z. Yi, R. Wu, X. Lu, P. Ma, and Y. Ma. Hau-net: Hybrid cnn-transformer for breast ultrasound image segmentation. *Biomedical Signal Processing and Control*, 87:105427, 2024. 2

[41] Y. Zhang, X. Wang, H. Chen, and W. Zhu. Adaptive disentangled transformer for sequential recommendation. In *Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining*, pages 3434–3445, 2023. 3

[42] X. Zhao, P. Zhang, F. Song, C. Ma, G. Fan, Y. Sun, Y. Feng, and G. Zhang. Prior attention network for multi-lesion segmentation in medical images. *IEEE Transactions on Medical Imaging*, 41(12):3812–3823, 2022. 3

[43] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. 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] X. Zhu, W. Zhou, and H. Li. Improving deep neural network sparsity through decorrelation regularization. In *Ijcai*, volume 8, pages 79–2, 2018. 3