2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 979-8-3503-3748-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/BIBM58861.2023.10386035

IMAN: An Iterative Mutual-Aid Network for Breast Lesion Segmentation on Multi-modal Ultrasound Images

Xiaozheng Xie
School of Comp. and Comm. Eng.
University of Science and Technology Beijing
Beijing, China
xiexiaozheng@ustb.edu.cn

Chen Chen
Hangzhou Innovation Institute
Beihang University
Hangzhou, China
zy1906702@buaa.edu.cn

Xuefeng Liu*
School of Comp. Sci. and Eng.
Beihang University
Zhongguancun Laboratory
Beijing, China
liu_xuefeng@buaa.edu.cn

Yong Wang
Chinese Academy of Medical Sciences and Peking Union Medical College
Beijing, China
drwangyong77@163.com

Rui Wang
School of Comp. and Comm. Eng.
University of Science and Technology Beijing
Beijing, China
wangrui@ustb.edu.cn

Jianwei Niu
School of Comp. Sci. and Eng.
Beihang University
Zhongguancun Laboratory
Beijing, China
niujianwei@buaa.edu.cn

***Abstract*—In the past decade, significant advancements have been made in utilizing deep learning for breast lesion segmentation. Recently, researchers have increasingly focused on harnessing the power of multiple modalities, recognizing its potential for enhancing segmentation performance. We observe that in clinical practice, many radiologists often rely on two types of ultrasound images, namely ultrasound (US) and contrast-enhanced ultrasound (CEUS) data for diagnosis. This motivates us to propose a multi-modal segmentation network, called as IMAN (Iterative Mutual-Aid Network), based on these two modalities. The architecture of IMAN adopts a novel hourglass shape, featuring two branches connected by an 'X' pathway. One branch is dedicated to processing CEUS data, while the other branch handles US data. Each branch generates segmentation results specific to its respective modality. The 'X' pathway, realized by a margin mask generator module, serves as a bridge between these branches by forcing the segmentation results from one branch as additional input to the other. This head-to-tail pathway effectively facilitates mutual aid between the two modalities. In addition, we propose an iterative training policy during the training process to fully exploit the information from both US and CEUS data. Experimental results on a Breast-US-CEUS dataset comprising 169 samples demonstrate the effectiveness of IMAN, achieving Dice Similarity Coefficient of 83.96% and 81.16% for US images and CEUS videos, respectively. These scores surpass those obtained by many state-of-the-art segmentation methods. Furthermore, IMAN exhibits robust generalization capabilities across different segmentation structures.**

***Index Terms*—breast lesion segmentation, multi-modal, ultrasound, contrast-enhanced ultrasound, iterative training**

I. Introduction

Breast lesion segmentation plays a crucial role in aiding the diagnosis and treatment of breast diseases, as accurate segmentation enables radiologists to conduct more precise disease analysis and treatment planning. In recent years, various breast lesion segmentation methods based on deep learning have been developed. These methods have demonstrated their ability to generate good segmentation results on different modalities, including ultrasound (US) [1]–[3], contrast-enhanced ultrasound (CEUS) [4], and mammograms [5].

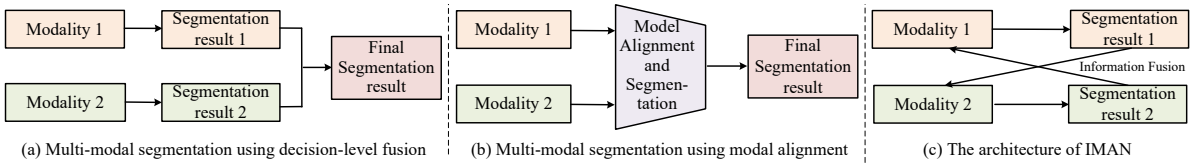
However, most of the aforementioned segmentation methods primarily focus on a single modality. This restriction often hampers the segmentation performance due to the low quality and limited information available from that specific modality. For instance, ultrasound is one of the most popular modalities utilized for deep learning models due to its low cost and widespread availability, but are frequently afflicted by issues such as speckle noise and low contrast, resulting in suboptimal segmentation outcomes for lesion boundaries. Although researchers have proposed various strategies including forcing models to focus more on boundary information in the segmentation process [6], adopting multi-task learning approaches [5], [7], the improvement is often limited.

Regarding the data modality for breast lesion segmentation, we observe that in clinical practice, many radiologists collect US and CEUS data simultaneously. US primarily displays the morphological information of lesions and provides a relatively clear visualization of the lesion margin. In contrast, CEUS offers a clear view of the infiltrated areas of the lesion [8]. We can see that these two modalities complement each other As evidence, radiologists commonly observe and compare the differences in lesion morphology between these two modalities, as it contains valuable diagnostic information [9].

This motivates us to propose a segmentation network based on US and CEUS. A straightforward approach is to directly

*Corresponding author
Xiaozheng Xie and Chen Chen contribute equally to this work.

Authorized licensed use limited to: Univ of Science and Tech Beijing. Downloaded on January 20,2024 at 08:23:46 UTC from IEEE Xplore. Restrictions apply.
979-8-3503-3748-8/23/$31.00 ©2023 IEEE



(a) Multi-modal segmentation using decision-level fusion (b) Multi-modal segmentation using modal alignment (c) The architecture of IMAN

Fig. 1. Different multi-modal segmentation structures.

utilize conventional multi-modal deep learning models for segmentation. However, this approach would fail in our scenario for the following reasons. Firstly, most current multi-modal segmentation methods adopt decision-level fusion (shown in Fig. 1(a)), by which different modalities are first segmented separately, then their results are fused using some averaging or majority voting methods [10]. This approach often overlooks the spatial alignment of multi-modal data and cannot capture the differences of the same feature across different modalities. In contrast, radiologists typically compare the difference between US and CEUS, as these two modalities are spatially aligned during the data acquisition process. Secondly, many segmentation methods using model alignment or fusion only generate segmentation results for one main modality [4], [11] (shown in Fig. 1(b)). However, in our scenario, segmentation results from both modalities are crucial, as differences between these two kinds of segmentation results are highly informative for radiologists in making decisions [9].

In this paper, we propose a novel multi-modal segmentation network called IMAN (Iterative Mutual-Aid Network) based on US and CEUS images. The architecture of IMAN is a unique hourglass shape (shown in Fig. 1(c)), consisting of two branches interconnected by an 'X' pathway. Each branch is dedicated to processing data of one modality, and generates segmentation results specific to its respective modality. Realized by a margin mask generator (MMG) module, the 'X' pathway serves as a bridge between these branches by forcing the segmentation results from one branch as additional input to the other. We also design an iterative training policy (ITP) for the training process, with which the margin information of the two modalities interacts iteratively during the learning process, leading to further improvements in segmentation.

We conduct extensive experiments on our self-built Breast-US-CEUS dataset. Experimental results demonstrate the superior performance of IMAN, achieving a Dice Similarity Coefficient (DSC) score of 83.96% on US images and 81.16% on CEUS videos. These results surpass several renowned segmentation frameworks and popular US/CEUS based segmentation methods. Experimental results also demonstrate the versatility and adaptability of IMAN by using different segmentation frameworks. The contributions of this paper can be summarized as follows:

- We propose IMAN, an iterative mutual-aid network for breast lesion segmentation using US and CEUS modalities. IMAN adopts a novel hourglass shape to achieve spatial feature alignment and mutual segmentation in multi-modal data. This innovative architecture allows for effective integration of information from both modalities.
- We introduce a MMG module that emphasizes the lesion margin from segmentation results. This module, combined with the proposed ITP, continually refines the generated margin masks and the segmentation results. By focusing on the margin, IMAN is able to capture important details and accurately delineate lesion boundaries.
- Through extensive experiments on a Breast-US-CEUS dataset, we demonstrate the effectiveness of IMAN compared to other popular segmentation methods. Meanwhile, IMAN also exhibits good generalization when integrated into various segmentation frameworks. These results validate the effectiveness and robustness of IMAN.

II. Related work

A. Breast Lesion Segmentation using Single-modal Images

Recently, various deep learning-based methods have been proposed for breast lesion segmentation across different modalities, such as US [1], [3], CEUS [4], mammogram [12], and magnetic resonance imaging (MRI) [13]. Among these modalities, US is most widely used due to its advantages of low cost and non-invasiveness. For example, Chen et al. [1] develop an AAU-net with self-attention and spatial-attention blocks for tumor segmentation in US images. Byra et al. [14] adopt the improved U-Net with a selective kernel for breast tumor segmentation. Some methods explore and exploit background-salient representation [2] or lesion boundary [6], which also achieve superior segmentation performance and robustness in US images.

Besides US, some studies also focus on other modalities. For instance, Singh et al. [12] apply conditional GAN to segment breast tumors within the region of interest in mammograms, achieving high-quality segmentation results. Min et al. [15] propose a multi-task method for mass detection and segmentation on pseudo-color mammograms. For better breast cancer segmentation in dynamic contrast-enhanced MRI (DCE-MRI), Lv et al. [13] propose a hybrid scheme by incorporating pharmacokinetics prior and feature refinement.

Overall, the aforementioned methods demonstrate good lesion segmentation performance on different modalities. However, the effectiveness of these approaches is often constrained by the limited information available from a single modality.

B. Breast Lesion Segmentation using Multi-modal Images

Recently, many deep learning-based multi-modal segmentation methods are employed in breast lesions [4], [16]. For instance, an inter-modality information interaction network is proposed for breast tumor segmentation in 3D

Authorized licensed use limited to: Univ of Science and Tech Beijing. Downloaded on January 20,2024 at 08:23:46 UTC from IEEE Xplore. Restrictions apply.

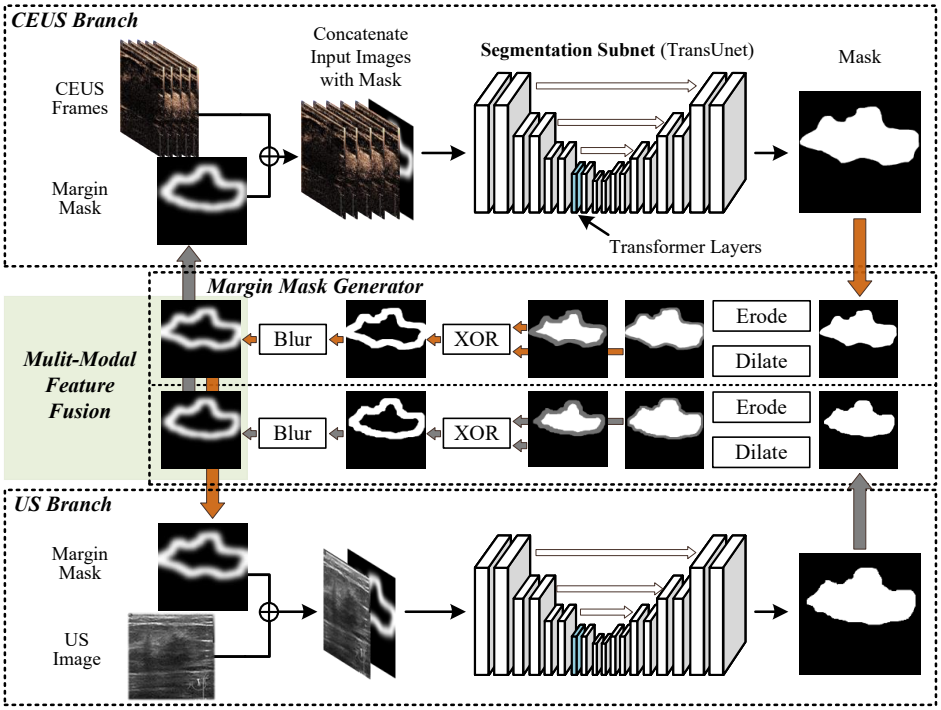


Fig. 2. The overall architecture of IMAN. It mainly contains two branches to process CEUS and US data separately. In between the two branches is an 'X' pathway (the margin mask generator (MMG) module) connecting them together, by which the segmentation results of one branch are injected into the input of another branch. Additionally, an iterative training policy (ITP) is adopted to continuously improve the quality of margin masks and segmentation results.

multi-modal MRI. This network incorporates a bi-directional request-supply information interaction module between different modalities [11]. In [16], an attention block is introduced to guide the information selection process, showing enhanced segmentation performance on T2-weighted and T1-weighted contrast-enhanced magnetic resonance images. In addition, there are other methods that combine US images with mammograms [17] or using multi-modal US images [4] for breast lesion segmentation.

However, the above multi-modal segmentation models would fail in our scenario. Firstly, most current multi-modal segmentation methods mainly rely on decision-level fusion, which ignores the spatial alignment of multi-modal data and overlooks differences in the same feature across modalities. Secondly, some modal alignment and feature fusion-based multi-modal segmentation methods only provide segmentation results for one main modality. However, in our scenario, the results from these two modalities are crucial.

III. Methods

The architecture of IMAN is shown in Fig. 2. We can see that IMAN has a unique hourglass shape, with two branches (one for CEUS and the other for US) interconnected by an 'X' pathway (the margin-mask generator).

A. Backbone Structure

The backbone of IMAN is a dual-branch segmentation network, each branch is a segmentation subnet designed for a single modality, and both subnets share the same structure. We use the TransUNet [18] structure as the segmentation subnet as it enables the learning of detailed high-resolution spatial information and global context information thus generally having high segmentation performance.

Specifically, TransUNet consists of a CNN-Transformer hybrid encoder and a cascaded upsampler. The encoder part starts by using ResNet50 [19] to extract features from the input images. These features are then passed through the patch embedding and transformer encoder layers, the latter composes of L layers of Multihead Self-Attention (MSA) and Multi-Layer Perceptron (MLP) blocks. In the upsampling process, multiple upsampling steps are designed to decode the hidden features encoded by Transformers and generate the final segmentation results. For more details of TransUNet, we refer readers to Chen et al. [18]. In this way, two branches generate segmentation masks for two modalities, respectively.

For the US branch, the input size is $C \times H \times W$, where C, H and W are the image channel, height and width, respectively. In the CEUS branch, the input size of the segmentation model is set as $T \times H \times W$, where T represents the frame number extracted from the CEUS video. During the training process, T frames are stacked in the channel dimension and as the input. Note that the network weights of these two branches are not shared during the training process.

B. Margin Mask Generator

Lesion margins in US and CEUS generally show different features. In CEUS, there are noticeable infiltrated areas, but the boundary of the lesion appears blurred. In US, the lesion

Authorized licensed use limited to: Univ of Science and Tech Beijing. Downloaded on January 20,2024 at 08:23:46 UTC from IEEE Xplore. Restrictions apply.

boundary is relatively clear, but the infiltrated areas are not easily observed. To fully utilize these two modalities, we develop a margin mask generator (MMG) module to extract the margin information from the segmentation result of one branch and use it in the segmentation process in another branch. The margin information is represented as the margin mask areas (abbreviated as margin masks hereinafter).

To generate margin masks, MMG implements a series of operations including erode, dilate, exclusive or operation (XOR), and blur (i.e., filtering), on the segmentation results. The erode operation aims to reduce the inner area of the segmentation results, while the dilate operation enlarges the external area. By applying the XOR operation, the model is encouraged to focus more on the surrounding areas of the lesion margin. Additionally, the blur operation allows for the conversion of pixel values in the margin masks from hard labels (i.e., 0 or 1) to soft ones, which facilitates the probabilistic fitting of the pixel values during the segmentation process.

Then the margin masks generated from the output of each branch is injected into the input of another one. This exchange process allows for the utilization of margin information learned from US images to assist the segmentation of lesion margins in the CEUS branch, and the US branch can also benefit from the infiltrated region information extracted from CEUS images.

C. Iterative Training Policy

As the margin masks generated from each branch are applied to another segmentation branch within IMAN, this training process forms a loop between the US and CEUS branches. In order to improve the performance of lesion segmentation, we dynamically select the margin masks with higher quality for the next training loop. To achieve this, we propose an iterative training policy (ITP) that enforces the entire training process of IMAN.

In the ITP, the training process of IMAN consists of multiple training loops, each composed of a few epochs. To achieve convergence in the training of each segmentation loop, the number of epochs contained within each loop is not fixed. Once the segmentation subnet reaches convergence at the l^{th} loop, the margin masks with higher quality are generated for each modality. These margin masks are then applied in the $(l+1)^{th}$ loop. This iterative approach ensures the lesion segmentation results and margin masks are continuously refined based on the information learned from previous training loops. The complete training process of IMAN using ITP is described in Algorithm 1, where Net and w are the segmentation network and its training weights, respectively.

With ITP, IMAN is trained in US and CEUS modalities iteratively. The final segmentation performance of each modality is constrained by the Dice loss defined as:

$$L_{dice} = 1 - \frac{2\sum_1^N p_i y_i + \varepsilon}{\sum_1^N p_i + \sum_1^N y_i + \varepsilon}, \tag{1}$$

where N represents the number of pixels in each image, and p_i and y_i denote the predicted probability and ground truth label, respectively, for the i^{th} pixel in the image. ε is a smooth factor and set as 1e-5. During the inference phase, the best model obtained from the training process is first used to generate high-quality margin masks which are then utilized to guide the generation of accurate lesion segmentation results.

Algorithm 1 The training process of IMAN

Input: Multi-modal images $input_{us}$, $input_{ceus}$
Initialization:
  $loop \leftarrow 0, epoch \leftarrow 0$
  $w_{us} \leftarrow w_{pretrained}, w_{ceus} \leftarrow w_{pretrained}$
repeat
  repeat
    if $loop = 0$ **then**
      $mask_{us} \leftarrow Net\{w_{us}, input_{us}\}$
      $mask_{ceus} \leftarrow Net\{w_{ceus}, input_{ceus}\}$
    else
      $mask_{us} \leftarrow Net\{w_{us}, input_{us}, margin_{ceus}\}$
      $mask_{ceus} \leftarrow Net\{w_{ceus}, input_{ceus}, margin_{us}\}$
    end if
    $epoch \leftarrow epoch + 1$
  until
    $\Delta L_{Dice} < \varepsilon$,
    $\Delta L_{Dice} \leftarrow L_{Dice_{epoch}} - L_{Dice_{(epoch-1)}}$
  update margin masks when reach convergence
  $margin_{us} \leftarrow MMG\{mask_{us}\}$
  $margin_{ceus} \leftarrow MMG\{mask_{ceus}\}$
  $epoch \leftarrow 0$
  $loop \leftarrow loop + 1$
  learning rate decay
until reach the max number of loop
Output: Lesion masks $mask_{us}$, $mask_{ceus}$

IV. Experiments

A. Dataset and Experimental Setup

Dataset: As there is no publicly available US-CEUS dataset, we evaluate the segmentation performance of IMAN using our self-constructed Breast-US-CEUS dataset. The dataset consists of 169 samples collected from 165 patients at the Cancer Hospital and Institute, Peking Union Medical College, and Chinese Academy of Medical Science. Each sample in the dataset includes one US image and its corresponding CEUS video. Both modalities are acquired simultaneously using a Philips EPIQ5 machine. The image sizes of the US image and frames in the CEUS videos are 460×560 pixels. Furthermore, each US-CEUS sample in the dataset has been labeled with a dual-mode mask by two experienced radiologists. This means that for each US image or a complete CEUS video, there is a corresponding segmentation mask. Note that the study protocol has been approved by the ethics committee of the participating hospital.

Implementation details: The dataset is randomly divided into the training, validation, and test sets, with a split ratio of 6:2:2. The US-based segmentation branch of IMAN accepts a single US image as input, while the CEUS-based branch incorporates 5 frames extracted from each CEUS video. Moreover, the input US and CEUS images are resized to a resolution of 320×320 pixels. Data augmentation techniques, such as random crop, horizontal flip, and rotation are also applied. Our experiments are implemented using the PyTorch framework.

Authorized licensed use limited to: Univ of Science and Tech Beijing. Downloaded on January 20,2024 at 08:23:46 UTC from IEEE Xplore. Restrictions apply.

The entire training process consists of 10 loops, each with a batch size of 16, and lasts for 300 epochs per loop. In particular, once reached convergence, the training process is early stopped in each loop. The learning rate is initialized to 0.0001 and then gradually reduced to a minimum learning rate of 5e-5 using the cosine annealing learning method. To ensure reproducibility and consistency of results, a carefully designed random seed is used. The model parameters of IMAN are pre-trained on the ImageNet dataset. All experiments are conducted on one NVIDIA Tesla V100-32G GPU. Our source code is available on https://github.com/xxzcs/IMAN.

Evaluation metrics: We choose Dice Similarity Coefficient (DSC) and Hausdorff distance (HD) as metrics to compare the performance of different methods. In particular, DSC measures the similarity of two segmentation sets, while Hausdorff distance (measured in millimeter) depicts the distance between the two boundary point sets of segmentation areas. Thus high DSC and low HD values generally mean better segmentation performance. The DSC and HD between the predicted results P and ground truth Y can be computed as follows:

$$DSC = \frac{2\left|Y \bigcap P\right|}{|Y| + |P|}, \tag{2}$$

$$HD = \max\left\{h(Y,P), h(P,Y)\right\}, \tag{3}$$

$$h(Y,P) = \max_{y\in Y}\min_{p\in P}\left\|y-p\right\|, \tag{4}$$

$$h(P,Y) = \max_{p\in P}\min_{y\in Y}\left\|p-y\right\|, \tag{5}$$

where p and y are the point sets of the predicted segmentation area P and the ground truth area Y, respectively. $h\left(\cdot,\cdot\right)$ is the one-way Hausdorff distance from the first to the second set, and $\left\|\cdot\right\|$ is the Euclidean distance.

B. Comparison With Other Segmentation Methods

We compare IMAN with six famous segmentation frameworks (i.e., UNet [20], FPN [21], UNet++ [22], DeepLabV3 [23], TransUNet [18] and SwinUnet [24]) and four segmentation methods that specifically designed for US or CEUS modalities (i.e., AAU-net [1], NU-net [3], CEUSegNet [4] and DpRAN [25]). These six segmentation frameworks are verified in US and CEUS modalities separately, in which multi-modal information is not utilized. Among the other four comparison methods, AAU-net and NU-net are specifically designed for US images, while CEUSegNet and DpRAN are designed for CEUS ones. In particular, even the segmentation process of CEUSegNet combines the information from US images, it only generates segmentation results for CEUS modality. Thus we also verify the segmentation performance on the corresponding modalities for these four methods.

The quantitative results of these methods are listed in Table I, and the best performance in each metric is shown in bold. We can see that IMAN significantly outperforms all comparison methods in almost all metrics, which achieves the DSC of 0.8396 and 0.8116, and the HD of 6.1450 and 6.8228 in US and CEUS modalities, respectively. When compared with these six famous segmentation networks, the DSC of IMAN is improved ranging from 2.26% to 6.51%, and from 4.27% to 7.49% in US and CEUS modalities, respectively. When compared with the methods specifically designed for US/CEUS images, IMAN also shows great advantages. In US modality, IMAN improves the DSC by 2.62% when compared with the suboptimal method NU-net, and the HD of IMAN is decreased by 0.4803. In CEUS modality, when compared with the suboptimal method CEUSegNet, IMAN improves the DSC by 2.92% and decreases the HD by 0.4939. To summarize, all comparison results demonstrate the effectiveness of IMAN in fusing multi-modal information. The advantages compared to CEUSegNet, which combines information from US, further underscore the effectiveness of the structure design of IMAN.

TABLE I
SEGMENTATION PERFORMANCE COMPARISON BETWEEN IMAN AND OTHER SEGMENTATION METHODS ON OUR BREAST-US-CEUS DATASET.

Method	US		CEUS	
	DSC↑	HD ↓	DSC↑	HD ↓
UNet [20]	0.7805	6.4727	0.7448	7.2023
FPN [21]	0.7893	6.4022	0.7428	6.8603
UNet++ [22]	0.7745	6.7443	0.7367	7.9155
DeepLabV3 [23]	0.7951	6.5712	0.7498	6.8880
TransUNet [18]	0.8033	**5.9172**	0.7689	7.2773
SwinUNet [24]	0.8170	6.6698	0.7590	7.8797
AAU-net [1]	0.7972	6.9505	—	—
NU-net [3]	0.8134	6.6253	—	—
DpRAN [25]	—	—	0.7767	7.3785
CEUSegNet [4]	—	—	0.7824	7.3167
IMAN(Ours)	**0.8396**	6.1450	**0.8116**	**6.8228**

The qualitative comparison results are shown in Fig. 3, where four pairs of US-CEUS samples are chosen for visualization. In Fig. 3, the top four rows are the segmentation results of US modality, and the bottom four are CEUS ones. For clearer presentation, the ground truth masks or predicted segmentation results are overridden in the original images. In particular, the middle frame in each CEUS video is chosen to show the segmentation results. It can be seen that: 1) For small lesions or lesions with irregular boundaries, segmentation methods like UNet, FPN, UNet++ and DeepLabV3 are more likely to over-segment or under-segment breast lesions (e.g., the first, second and last rows in both these two modalities). 2) For US modality, segmentation results of TransUNet and SwinUNet are generally smoother and insufficient to reflect the specificity of lesion boundary, while the results of AAU-net and NU-net can reflect the lesion specificity but sometimes with false positive predictions. 3) Since lesion areas are generally difficult to distinguish when contrast agents are injected into blood vessels, segmentation in CEUS modality is more difficult than that in US. Thus most methods, in particular the DpRAN and CEUSegNet, generate segmentation results with coarse lesion boundaries and shapes. 4) By letting multi-modal information assist each other, IMAN generates segmentation results more accurately in both US and CEUS modalities, in which the location and shape information of lesions can

Authorized licensed use limited to: Univ of Science and Tech Beijing. Downloaded on January 20,2024 at 08:23:46 UTC from IEEE Xplore. Restrictions apply.

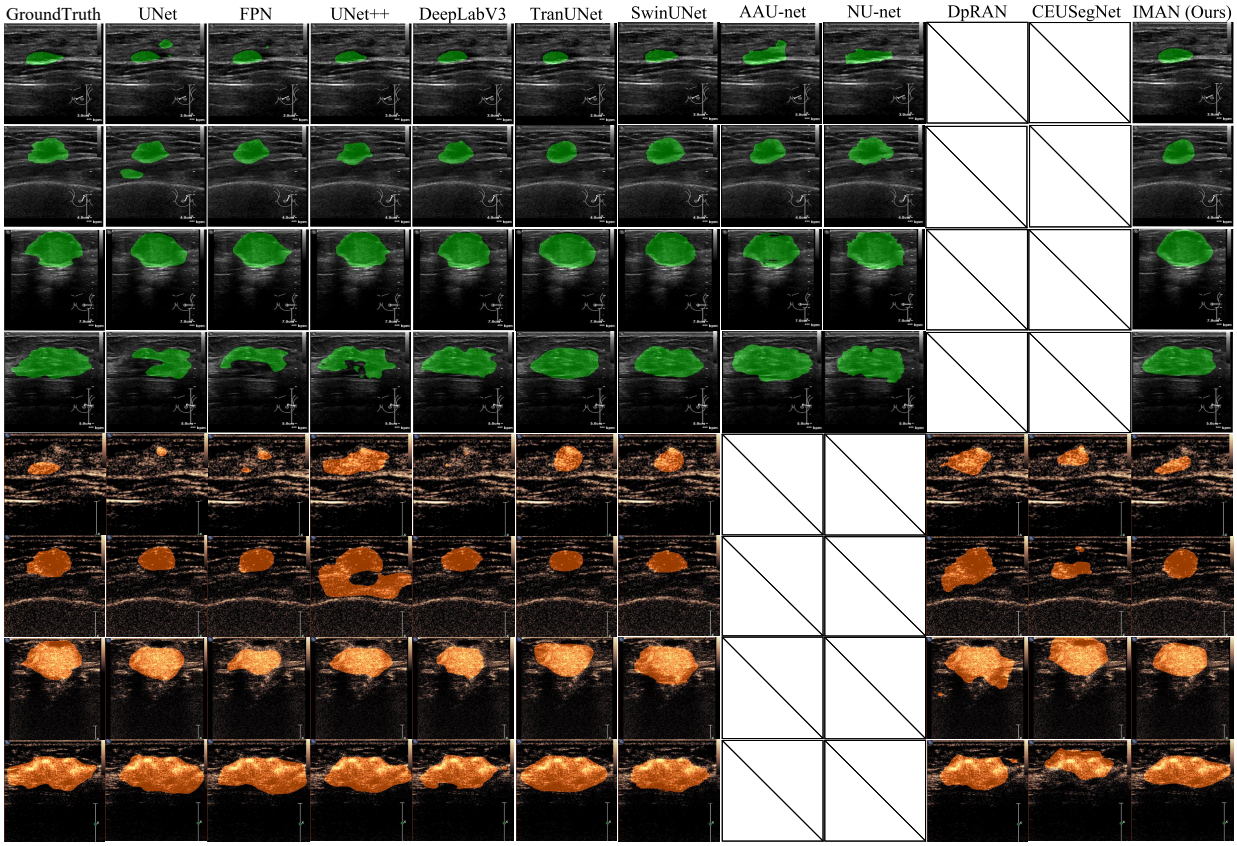


Fig. 3. Visualization of segmentation results of IMAN and other comparison methods. The top four rows are the results of US modality, while the bottom four are the CEUS ones.

be preserved. This again validates the effectiveness of the proposed hourglass-shaped information fusion structure.

C. Ablation Study

Effectiveness of different modules: We first investigate the efficacy of different modules in IMAN (i.e., the MMG, ITP, and modal fusion modules). It should be noted that MMG and ITP modules are conducted within the same modality when the modal fusion module is not adopted. Table II presents the quantitative results for these various conditions along with the baseline method (i.e., TransUNet), and the best performance in each metric is shown in bold. It is apparent that incorporating MMG module significantly enhances segmentation performance, leading to 2.35% and 2.01% increases in DSC, as well as the decrease of 0.0954 and 0.4469 in HD in US and CEUS modalities, respectively. This finding underscores the importance of utilizing lesion margin information during the segmentation process. Additionally, the integration of the ITP module with MMG ensures iterative updates to the generated margin masks and segmentation results throughout the training process, yielding the DSC of 0.8313 and 0.8033 in US and CEUS modalities, respectively. Finally, when the modal fusion module is added, significant enhancements have been achieved across most metrics, resulting in 3.63% and 4.27% increase in DSC for US and CEUS modalities, respectively. The results also demonstrate that, by combining all these three modules, IMAN can make full use of the difference between different modalities, which highly boosts the segmentation performance.

TABLE II
ABLATION STUDY OF THE MODULES IN IMAN.

MMG	ITP	Modal Fusion	US		CEUS	
			DSC↑	HD ↓	DSC↑	HD ↓
			0.8033	5.9172	0.7689	7.2733
✓			0.8268	**5.8218**	0.7890	6.8264
✓	✓		0.8313	6.1622	0.8033	6.8947
✓	✓	✓	**0.8396**	6.1450	**0.8116**	**6.8228**

TABLE III
SEGMENTATION PERFORMANCE ON CEUS VIDEO WHEN DIFFERENT NUMBERS OF FRAMES ARE USED.

Frame Num	DSC ↑	HD ↓
1	0.7011	7.0034
3	0.7496	**6.6131**
5	**0.7689**	7.2773
7	0.7544	7.2931
9	0.7543	7.2010

The impact of frame numbers in CEUS videos: To capture the lesion characteristics in infiltration areas over time, we extract various frames from CEUS videos and incorporate

Authorized licensed use limited to: Univ of Science and Tech Beijing. Downloaded on January 20,2024 at 08:23:46 UTC from IEEE Xplore. Restrictions apply.

them as input to the segmentation subnet. To examine the impact of the number of frames from CEUS videos on segmentation performance, we perform experiments with different frame numbers. Specifically, we focus on the segmentation of a single CEUS modality without any additional modules. We select the middle frame of the video as the starting point and extract frames from both sides of the middle one with a fixed time step (i.e., 3 frames). In particular, we test frame numbers 1, 3, 5, 7, and 9, and the results are presented in Table III. It is observed that there may be a slightly higher HD value, while the DSC is optimal when the frame number is set to $T = 5$. This setting is utilized in all our experiments.

Parameter ablation in MMG module: As mentioned in subsection III-B, the generation of margin masks from segmentation results involves the use of erode, dilate, and filtering operations. As the choice of kernel sizes for erode and dilate operations, as well as the selection of filters can impact the quality of the margin masks and the final segmentation results, we conduct experiments by combining different kernel sizes and filters. Specifically, kernel sizes of 50, 75, 100, and 125 are combined with mean filters, Gaussian filters, and no filter (referred to as 'None'), respectively. Moreover, since lesion inflation areas are generally larger in CEUS compared with that in US modality, the condition of kernel size of 125 is only experimented in the CEUS. The results for these conditions are presented in Table IV, where the first row represents the segmentation performance of the baseline method without introducing MMG module.

From Table IV, we can see that most combinations of kernel sizes and filters contribute to improvements in segmentation performance to some extent. Additionally, optimal performance is achieved with different settings for different modalities. For the US modality, the combination of a Gaussian filter and a kernel size of 75 achieves the highest DSC and the lowest HD. This configuration results in a 2.35% improvement in DSC and a decrease of 0.0954 in HD when compared with the baseline method. In contrast, for the CEUS modality, the best segmentation performance is attained when using a mean filter and a kernel size of 100. This combination improves the DSC by 2.01% and reduces the HD by 0.4509. The reason behind this may be attributed to the fact that in CEUS, the lesion boundary is often connected to blood vessels and therefore a larger kernel size is required to encompass the expanded areas of the lesion to achieve accurate segmentation.

TABLE IV
PERFORMANCE AT DIFFERENT PARAMETER SETTINGS IN MMG MODULE.

Kernel Size	Filter	US		CEUS	
		DSC ↑	HD ↓	DSC ↑	HD ↓
-	-	0.8033	5.9172	0.7689	7.2773
50	None	0.8132	6.1863	0.7493	7.0269
50	mean	0.7914	6.3364	0.7651	6.9656
50	gaussian	0.7728	6.2806	0.7566	7.5291
75	None	0.8150	5.6159	0.7324	**6.5209**
75	mean	0.8227	6.0159	0.7638	6.9645
75	gaussian	**0.8268**	**5.8218**	0.7566	7.0210
100	None	0.8177	6.3094	0.7625	6.6749
100	mean	0.8026	6.3260	**0.7890**	6.8264
100	gaussian	0.8165	5.8292	0.7688	7.2791
125	None	-	-	0.7427	6.8494
125	mean	-	-	0.7710	6.6905
125	gaussian	-	-	0.7614	6.8947

TABLE V
GENERALIZATION OF IMAN ON OTHER SEGMENTATION FRAMEWORKS.

Backbone	MMG	ITP	Modal Fusion	US		CEUS	
				DSC↑	HD ↓	DSC↑	HD ↓
FPN				0.7903	6.4022	0.7428	**6.8603**
	✓			0.7847	**0.0451**	0.7508	6.8752
	✓	✓		0.7807	7.6068	0.7570	7.7851
	✓	✓	✓	**0.8014**	7.1204	**0.7648**	8.5213
UNet				0.7805	**6.4727**	0.7448	7.2023
	✓			0.7825	6.8590	0.7249	6.9511
	✓	✓		0.7837	6.4904	0.7430	7.3937
	✓	✓	✓	**0.8000**	6.4964	**0.7929**	**6.9091**
UNet++				0.7745	6.7443	0.7367	7.9155
	✓			0.7946	6.5573	0.7344	7.0813
	✓	✓		0.8070	6.2608	0.7802	7.5847
	✓	✓	✓	**0.8201**	**6.2260**	**0.8047**	**6.6516**
DeepLabV3				0.7951	6.5712	0.7498	6.8880
	✓			0.7890	6.3626	0.7354	7.2403
	✓	✓		0.7876	**6.2701**	0.7872	6.7734
	✓	✓	✓	**0.8183**	6.3700	**0.7997**	**6.5589**
SwinUNet				0.8170	6.6698	0.7590	7.8797
	✓			0.8129	6.6259	0.7380	7.9762
	✓	✓		0.8210	6.6056	0.7758	7.6915
	✓	✓	✓	**0.8235**	**6.5955**	**0.7822**	**7.3351**

D. Robustness of IMAN on Other Segmentation Structures

To demonstrate the generalization of IMAN, we replace the original TransUNet module of IMAN with several popular segmentation architectures, including FPN, UNet, UNet++, DeepLabV3, and SwinUNet while keeping the training protocol the same as IMAN. The quantitative results of these different methods are presented in Table V, and the best metrics in each network structure are shown in bold. From the table, it can be observed that almost all segmentation methods achieve improved segmentation performance when incorporating all modules of IMAN. This finding underscores the robustness and generalization ability of IMAN.

More specifically, when the MMG module is adopted, most U-Net based methods perform better than its baseline method in US modality. The improvements of DSC ranged from 0.20% to 2.01%, and the decreases of HD ranged from 0.0439 to 0.3571. The reason behind this may be that the U-shaped structures can capture the margin information at incoming inputs more effectively as they generally fuse the information from shallow to deep layers. Moreover, there are more performance improvements in the US modality compared to the CEUS one. This discrepancy can be attributed to the fact that lesion margins are often ambiguous in CEUS images, and directly incorporating margin information once may not yield obvious improvements as it does in the US modality.

Authorized licensed use limited to: Univ of Science and Tech Beijing. Downloaded on January 20,2024 at 08:23:46 UTC from IEEE Xplore. Restrictions apply.

Additionally, when ITP module is added on top of MMG, there is a substantial performance improvement in almost all metrics for UNet++, DeepLabV3 and SwinUNet architectures. As margin masks generated by the MMG module are continuously updated by ITP, the lesion segmentation performance can also be improved. Finally, when exchanging margin masks of US and CEUS modalities, the DSC is the highest in all network structures on both modalities, which is improved from 0.65% to 4.56% on US images and from 2.40% to 6.80% on CEUS frames. This also demonstrates that the integration of multi-modal information helps the model capture the differences in lesion characteristics, thereby can highly improve the segmentation performance in both modalities.

V. Conclusion

In this paper, we propose IMAN, an hourglass-shaped iterative training framework for lesion segmentation in US and CEUS modalities. IMAN consists of two segmentation branches (i.e., the US branch and the CEUS one), and an 'X' pathway to exchange margin information extracted from segmentation rseults between the two branches. Besides, an iterative training policy is designed to continually refine the quality of margin masks and segmentation results. Experiments on a Breast-US-CEUS dataset show the superior performance of IMAN compared with other state-of-the-art methods, as well as its robustness and generalization ability. We hope that IMAN will inspire the framework design of future lesion segmentation on multi-modal images. In the future, we plan to conduct clinical testing to further validate the advantages demonstrated of IMAN.

Acknowledgment

This work was supported by China Postdoctoral Science Foundation [grant numbers 2023M730224, 2017M620683], National Natural Science Foundation of China [grant numbers 61976012, 62372028, 62372027], and Fundamental Research Funds for the Central Universities [grant number FRF-TP-22-051A1].

References

[1] G. Chen, L. Li, Y. Dai, J. Zhang, and M. H. Yap, "Aau-net: an adaptive attention u-net for breast lesions segmentation in ultrasound images," *IEEE Transactions on Medical Imaging*, vol. 42, no. 5, pp. 1289–1300, 2023.

[2] Z. Ning, S. Zhong, Q. Feng, W. Chen, and Y. Zhang, "Smu-net: Saliency-guided morphology-aware u-net for breast lesion segmentation in ultrasound image," *IEEE Transactions on Medical Imaging*, vol. 41, no. 2, pp. 476–490, 2022.

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

[4] Z. Meng, Y. Zhu, X. Fan, J. Tian, F. Nie, and K. Wang, "Ceusegnet: A cross-modality lesion segmentation network for contrast-enhanced ultrasound," in *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2022, pp. 1–5.

[5] J. Wang, Y. Zheng, J. Ma, X. Li, C. Wang, J. Gee *et al.*, "Information bottleneck-based interpretable multitask network for breast cancer classification and segmentation," *Medical Image Analysis*, vol. 83, p. 102687, 2023.

[6] R. Huang, M. Lin, H. Dou, Z. Lin, Q. Ying, X. Jia *et al.*, "Boundary-rendering network for breast lesion segmentation in ultrasound images," *Medical Image Analysis*, vol. 80, p. 102478, 2022.

[7] H. Jiao, X. Jiang, Z. Pang, X. Lin, Y. Huang, and L. Li, "Deep convolutional neural networks-based automatic breast segmentation and mass detection in dce-mri," *Computational and Mathematical Methods in Medicine*, vol. 2020, 2020.

[8] J. Folkman, "Role of angiogenesis in tumor growth and metastasis," in *Seminars in oncology*, vol. 29, no. 6. Elsevier, 2002, pp. 15–18.

[9] C. Chen, Y. Wang, J. Niu, X. Liu, Q. Li, and X. Gong, "Domain knowledge powered deep learning for breast cancer diagnosis based on contrast-enhanced ultrasound videos," *IEEE Transactions on Medical Imaging*, vol. 40, no. 9, pp. 2439–2451, 2021.

[10] T. Zhou, S. Ruan, and S. Canu, "A review: Deep learning for medical image segmentation using multi-modality fusion," *Array*, vol. 3, p. 100004, 2019.

[11] C. Peng, Y. Zhang, J. Zheng, B. Li, J. Shen, M. Li *et al.*, "Imiin: An inter-modality information interaction network for 3d multi-modal breast tumor segmentation," *Computerized Medical Imaging and Graphics*, vol. 95, p. 102021, 2022.

[12] V. K. Singh, H. A. Rashwan, S. Romani, F. Akram, N. Pandey, M. M. K. Sarker *et al.*, "Breast tumor segmentation and shape classification in mammograms using generative adversarial and convolutional neural network," *Expert Systems with Applications*, vol. 139, p. 112855, 2020.

[13] T. Lv, Y. Wu, Y. Wang, Y. Liu, L. Li, C. Deng *et al.*, "A hybrid hemodynamic knowledge-powered and feature reconstruction-guided scheme for breast cancer segmentation based on dce-mri," *Medical Image Analysis*, vol. 82, p. 102572, 2022.

[14] M. Byra, P. Jarosik, A. Szubert, M. Galperin, H. Ojeda-Fournier, L. Olson *et al.*, "Breast mass segmentation in ultrasound with selective kernel u-net convolutional neural network," *Biomedical Signal Processing and Control*, vol. 61, p. 102027, 2020.

[15] H. Min, D. Wilson, Y. Huang, S. Liu, S. Crozier, A. P. Bradley *et al.*, "Fully automatic computer-aided mass detection and segmentation via pseudo-color mammograms and mask r-cnn," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2020, pp. 1111–1115.

[16] C. Li, H. Sun, Z. Liu, M. Wang, H. Zheng, and S. Wang, "Learning cross-modal deep representations for multi-modal mr image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 57–65.

[17] K. Atrey, B. K. Singh, A. Roy, and N. K. Bodhey, "Real-time automated segmentation of breast lesions using cnn-based deep learning paradigm: Investigation on mammogram and ultrasound," *International Journal of Imaging Systems and Technology*, vol. 32, no. 4, pp. 1084–1100, 2022.

[18] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang *et al.*, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[19] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1492–1500.

[20] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.

[21] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2117–2125.

[22] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Transactions on Medical Imaging*, vol. 39, no. 6, pp. 1856–1867, 2019.

[23] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[24] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian *et al.*, "Swin-unet: Unet-like pure transformer for medical image segmentation," in *European Conference on Computer Vision*, 2022, pp. 205–218.

[25] P. Wan, H. Xue, C. Liu, F. Chen, W. Kong, and D. Zhang, "Dynamic perfusion representation and aggregation network for nodule segmentation using contrast-enhanced us," *IEEE Journal of Biomedical and Health Informatics*, 2023.

Authorized licensed use limited to: Univ of Science and Tech Beijing. Downloaded on January 20,2024 at 08:23:46 UTC from IEEE Xplore. Restrictions apply.