TaiChiNet: Negative-Positive Cross-Attention Network for Breast Lesion Segmentation in Ultrasound Images

Article *in* IEEE Journal of Biomedical and Health Informatics · January 2024

DOI: 10.1109/JBHI.2024.3352984

CITATIONS
10

READS
282

6 authors, including:

Jinting Wang
Southern Medical University
4 PUBLICATIONS **31** CITATIONS
SEE PROFILE

Jiafei Liang
Southern Adventist University
4 PUBLICATIONS **70** CITATIONS
SEE PROFILE

Yang Xiao
Huazhong University of Science and Technology
149 PUBLICATIONS **4,187** CITATIONS
SEE PROFILE

Fang Zhiwen
Huazhong University of Science and Technology
45 PUBLICATIONS **766** CITATIONS
SEE PROFILE

TaiChiNet: Negative-Positive Cross-Attention Network for Breast Lesion Segmentation in Ultrasound Images

Jinting Wang, Jiafei Liang, Yang Xiao, Joey Tianyi Zhou, Zhiwen Fang, Feng Yang

***Abstract*— Breast lesion segmentation in ultrasound images is essential for computer-aided breast-cancer diagnosis. To improve the segmentation performance, most approaches design sophisticated deep-learning models by mining the patterns of foreground lesions and normal backgrounds simultaneously or by unilaterally enhancing foreground lesions via various focal losses. However, the potential of normal backgrounds is underutilized, which could reduce false positives by compacting the feature representation of all normal backgrounds. From a novel viewpoint of bilateral enhancement, we propose a negative-positive cross-attention network to concentrate on normal backgrounds and foreground lesions, respectively. Derived from the complementing opposites of bipolarity in TaiChi, the network is denoted as TaiChiNet, which consists of the negative normal-background and positive foreground-lesion paths. To transmit the information across the two paths, a cross-attention module, a complementary MLP-head, and a complementary loss are built for deep-layer features, shallow-layer features, and mutual-learning supervision, separately. To the best of our knowledge, this is the first work to formulate breast lesion segmentation as a mutual supervision task from the foreground-lesion and normal-background views. Experimental results have demonstrated the effectiveness of TaiChiNet on two breast lesion segmentation datasets with a lightweight architecture. Furthermore, extensive experiments on the thyroid nodule segmentation and retinal optic cup/disc segmentation datasets indicate the application potential of TaiChiNet.**

This work was supported in part by the National Natural Science Foundation of China under Grant 62371219, 61771233 and 62271221, Guangdong Basic and Applied Basic Research Foundation under Grant No.2023A1515011260, Science and Technology Program of Guangzhou under Grant No. 202201011672, SERC Central Research Fund (Use-inspired Basic Research). (Corresponding authors: Zhiwen Fang and Feng Yang).

Jinting Wang, Jiafei Liang, Zhiwen Fang, and Feng Yang are with the School of Biomedical Engineering, Southern Medical University, Guangzhou 510515, China (E-mail: 3168010128@i.smu.edu.cn; garfield6@i.smu.edu.cn; fzw310@smu.edu.cn; yangf@smu.edu.cn). They are also with the Guangdong Provincial Key Laboratory of Medical Image Processing, Southern Medical University, Guangzhou 510515, China and the Guangdong Province Engineering Laboratory for Medical Imaging and Diagnostic Technology, Southern Medical University, Guangzhou 510515, China.

Yang Xiao is with the National Key Laboratory of Multispectral Information Intelligent Processing Technology, School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: Yang_Xiao@hust.edu.cn).

Joey Tianyi Zhou is with the Centre for Frontier AI Research (CFAR), Research Agency for Science, Technology, and Research (A*STAR), Singapore 138632; he is also with the Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), Singapore 138632 (e-mail: zhouty@cfar.a-star.edu.sg).

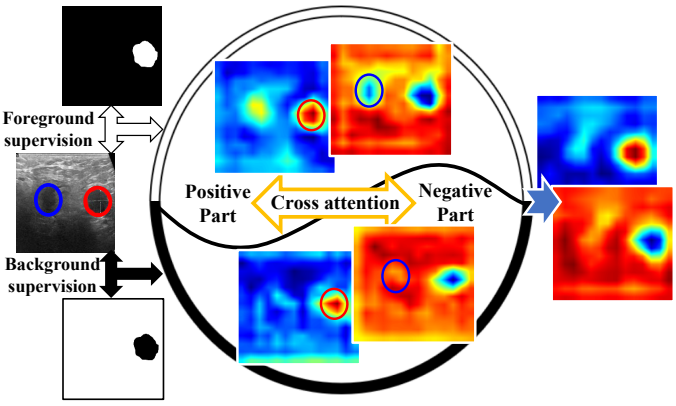

Fig. 1. The motivation of the negative-positive cross-attention network. To reduce the ambiguity produced by intricately comparable patterns between foreground lesions and normal backgrounds, separate pattern mining of the foreground and background would be essential. A lesion region is marked by the red circle, while a distracted region is indicated by the blue circle.

***Index Terms*— Negative-positive cross-attention, Breast lesion segmentation, Mutual learning, Ultrasound images.**

I. Introduction

BREAST cancer is one of the dreadful diseases that poses a grave danger to the health and lives of women [1]. Early detection of breast cancer can lower mortality by up to 20% [2], [3]. Automatic high-quality segmentation of breast lesions is a crucial step in computer-aided diagnosis (CAD) of breast cancer [4], and can be characterized as a pixel-level binary classification problem in which some pixels indicate lesions and others represent normal backgrounds. However, the segmentation performance is often weakened due to intricate ultrasonic patterns, speckle noise, and shadows.

With the goal of improving the segmentation performance, numerous researchers concentrate on developing an elegant model by simultaneously mining patterns of foreground lesions and normal backgrounds using global [5]–[7] or local [8]–[10] attention features. Due to the similarity of patterns between foreground lesions and normal backgrounds, the confusion in feature representation is hard to avoid. It often degrades the segmentation performance under the sensitivity attribute due to the class imbalance problem. To handle this problem, some approaches [11]–[15] pay more attention to foreground lesions by labeling the lesions as 1, and employing

various 1-label focal losses (e.g., dice loss). The methods that focus on foreground lesions unilaterally obtain an increase in true positives in the lesion foreground but also lead to an increase in false positives in the normal background, which includes distracted regions.

To alleviate the aforementioned issues, an alternate perspective is the bilateral enhancement of foreground lesions and normal background, followed by cross aggregation. Given supervision (e.g., dice coefficient loss) focusing on the 1-label category, as seen in Fig. 1, the foreground-lesion path and the normal-background path get features with varying degrees of bias. For clarity, the features offered by 1-label supervisions of foreground lesions and normal backgrounds are distinguished as the positive and negative parts, respectively. Specifically, the positive part of the foreground-lesion path can more effectively highlight the red-circled lesion region than the negative part of the normal-background path. However, a similar background region (circled in blue) close to the red-circled lesion also elicits a strong feature response. In contrast, the normal-background path compacts the feature of the blue-circled region with that of other backgrounds. Obviously, the diversified features provide complement potential across the channels of the two paths. This observation offers the opportunity to improve segmentation performance via cross-information interaction.

Following the motivation of a bilateral enhancement, we formulate breast lesion segmentation as a mutual supervision task between the negative normal-background path and the positive foreground-lesion path. To mimic the mutual supervision task [16], a negative-positive cross-attention network designated TaiChiNet,[1] is proposed via a cross-attention module and a complementary head. TaiChiNet consists of two U-Nets, which mine the patterns of foreground lesions and normal backgrounds, respectively. Between the deep layers of two U-Nets, a cross-attention module is designed to transfer the information across deep semantic channels. Following the decoded features of two U-Nets, a complementary head with multiple interactive MLPs is further introduced to refine the shallow features. Moreover, to improve the feature representation under mutual supervision, a complementary loss between the negative and positive paths is offered to deliver the bilateral complementary information on top of traditional segmentation constraints, where the dice loss focuses primarily on the unilateral enhancement. Finally, inspired by the efficacy of gradually difficulty-level learning [18] in various computer vision fields [19]–[21], we develop an easy-to-hard learning strategy to direct the model's attention to incrementally difficulty-level regions throughout the training phase of breast lesion segmentation.

The source code of this work is published online.[2]

The main contributions of this article are as follows.

• Aiming to improve segmentation performance by increasing true positives in the lesion foreground while restraining false positives in the normal background, a negative-positive cross-attention network is proposed for breast lesion segmentation. To the best of our knowledge, it is the first work of forming breast lesion segmentation as a mutual learning task via bilateral interaction between foreground lesions and normal backgrounds.

• In an effort to transmit information between foreground lesions and normal backgrounds in the mutual learning task, we design a cross-attention module, a complementary head, and a complementary loss for deep-layer features, shallow-layer features, and supervision, respectively. To simultaneously fuse deep-layer features and restrain information redundancy, we build a channel-to-spatial cross-attention module C^2-attention, which is different from the traditional cascaded channel and spatial attention.

• A pixel-level easy-to-hard learning strategy is provided to progressively improve the feature representation using the same training data and easy-to-hard supervision in order to resolve the varying segmentation difficulties among pixels within foreground lesions.

The remainder of this article is organized as follows. The related work is introduced in Sec. II. Then,

TaiChiNet and the objective function are detailed in Sec. III and IV, respectively. Experiments and discussions are conducted in Sec. V. Sec. VI concludes the whole paper.

II. Related work

Significant advances have been made in breast lesion segmentation during the last few decades. In this section, we first go through the recent techniques for segmenting breast lesions. Then, relevant research studies investigating the complementary information of the background are summarized.

A. Breast Lesion Segmentation Methods in Ultrasound Images

Breast lesion segmentation in ultrasound images is challenging due to the speckle noise, shadows, low contrast, ambiguous boundaries, and variances in lesion shape and size. It has evolved tremendously because of the extensive deployment of deep learning techniques. Numerous medical image segmentation tasks have demonstrated the benefits of including the attention mechanism in the networks. Some researchers [5]–[7] attempt to incorporate the attention mechanism into their model to improve breast lesion segmentation. By combining a boundary detection module, a spatial attention module, and a channel attention module, Xue et al. [5] build a global guidance network for breast lesion segmentation. A hybrid adaptive attention module is applied to replace the traditional convolution operation in [6], which consists of a channel attention block and a spatial attention block. To enhance the feature representation ability, a dual-attention that combines channel attention and lesion attention is proposed in [7]. It is undeniable that non-local features obtained by the attention mechanism are useful for the model to learn the discriminative features of the lesion. However, there are some distant pixels in the backgrounds that have an appearance similar to the

[1]TaiChi is an alternating principle of bipolarity (i.e., Yin and Yang) in Chinese philosophy [17]. Yin and Yang are not absolutes or opposing forces but rather complementing opposites [17]. TaiChiNet is the name of our network since it is inspired by the complementing opposites of negative-positive bipolarity.

[2]Code is available at https://github.com/beria-moon/TaiChiNet

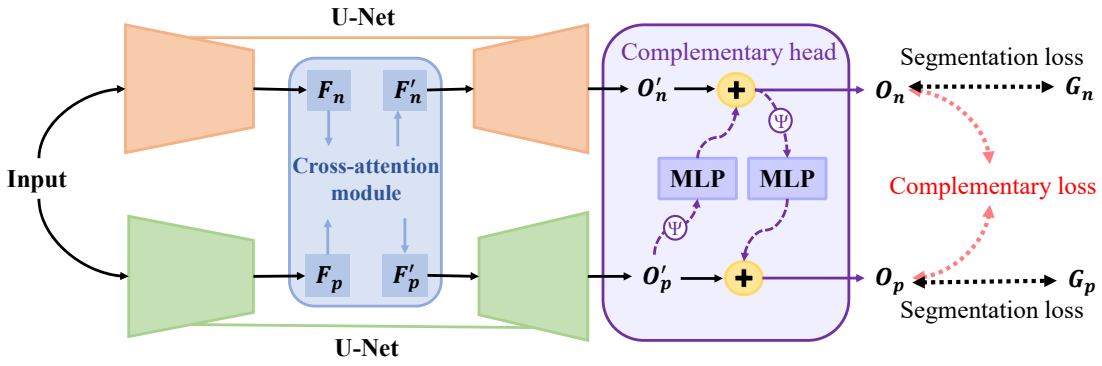

Fig. 2. Architecture of TaiChiNet, where "+" denotes element-wise addition and "Ψ" means complement operation as (12). TaiChiNet consists of two sub-network U-Nets, a cross-attention module, and a complementary head with two MLPs. Besides the traditional segmentation losses, a complementary loss is further designed for mutual supervision.

breast lesions, and incorporating these pixels in the long-term features may mislead the model to incorrectly classify these pixels as lesions.

Enlarging the receptive field by dilated convolutions, pooling operations, or fusing multi-level features also is a powerful tool to capture more lesion-related semantic features. Dilated convolutions are applied in the deeper layers to obtain a large respective field that benefits lesion separation from the background [8]. A dilated semantic segmentation network is proposed to segment the breast lesion by using progressively larger dilatations in each succeeding convolution layer in [10]. By fusing features from various scale levels, Li et al. [9] build a multi-scale fusion U-Net to obtain features with multiple receptive fields. Similarly, a feature-compression-pyramid network (FCP-Net) is proposed to integrate the multi-level features in [22]. Shareef et al. [23] apply two encoders to extract and fuse image context information at different scales. Enlarging the receptive field would help to capture more lesion-related semantic features. However, it only takes into account the inter-dependencies among spatial domains, and its capability will be limited when it deals with ambiguous boundaries.

The above-mentioned attempts have advanced performance, but they are still inadequate to address the confusion in the backgrounds brought by speckle noise or shadows because of the underutilized background information.

B. Foreground-background Integration Segmentation

In the field of video object segmentation, researchers have extensively investigated the use of background information to improve the accuracy of foreground segmentation. They have focused on resolving issues such as dynamic backgrounds, sudden illumination changes, and the influence of shadows [24]–[27]. Yang *et al.* [24] tackle inaccuracies in foreground-specific features by regularizing them using background attention maps generated through background-aware pooling [25], resulting in improved foreground discrimination. In the work conducted by [26], a generative adversarial network is trained using synthetic paired photo-realistic images. This network is capable of disentangling an image into its foreground and background components, therefore reducing background noise and enhancing the accuracy of foreground segmentation. In [27], a noteworthy strategy involved using an auto-encoder to approximate the background noise. This estimation was then used as pixel-wise uncertainty to adaptively alter the threshold for generating the foreground segmentation mask. However, such research is scarce in medical image segmentation. For the segmentation of magnetic resonance (MR) images, Sauvalle *et al.* [28] propose a multi-task network to predict both the background and foreground simultaneously. Subsequently, they used the background prediction to enhance the rough foreground in a multi-stage process. These efforts increase the accuracy of foreground segmentation by using background information to unilaterally enhance the representation of the foreground. Ning *et al.* [29] use a method called the coarse-to-fine technique to improve the representations of foreground and background. They do this by including low-level and high-level saliency maps in a human-in-the-loop fashion. Differently, our proposed method achieves bilateral improvement in the representation of both foreground and background by exploiting the complementary information contained in the original image in an end-to-end paradigm.

III. Negative-Positive Cross-Attention Network

To accurately predict pixel-level classifications, it is critical for breast lesion segmentation methods to highlight lesions while minimizing confusion with comparable background regions. Customarily, decent context guidance [5], [30] and focal loss [13], [31] are widely employed to emphasize the foreground lesions. However, both methods would introduce false positives due to the unilateral enhancement of foreground lesions, and an additional auxiliary model is required for the context guidance. In order to properly highlight lesions under the assistance of backgrounds, user interaction is introduced to offer the prior knowledge of the probable lesion regions and background regions [29]. However, the user interaction will increase the extra burden for users and diminish application potential.

Instead of user interaction, we propose a negative-positive cross-attention network to bilaterally enhance lesions and backgrounds. The network is denoted as TaiChiNet, which forms the breast lesion segmentation as a mutual-supervision task without a human-in-the-loop. While meeting the requirement of bilateral enhancement with mutual learning, two facts

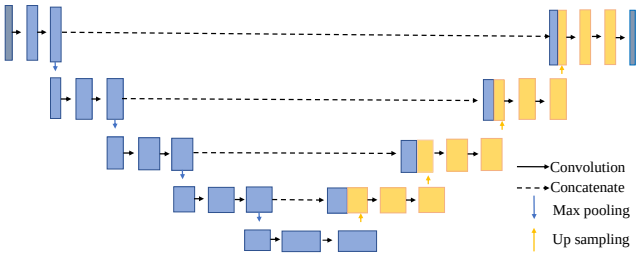

Fig. 3. Illustration of the sub-network U-Net, which is adopted as the backbone of TaiChiNet.

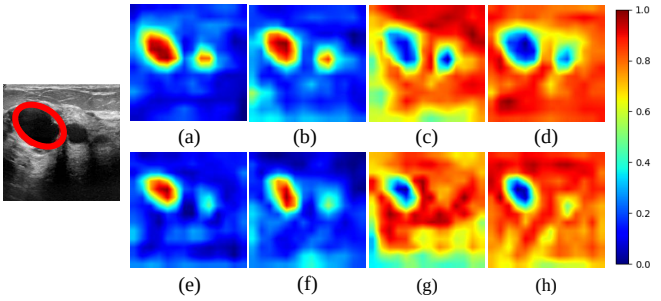

Fig. 4. Feature visualization of the negative and positive paths.(a)-(d) are four feature maps of the positive path; (e)-(h) are four feature maps of the negative path. The red-circled region is utilized to represent the ground truth. The higher value in the color bar means a stronger response.

are essential: 1) independent paths with the potential to unilaterally enhance lesions and background; and 2) hierarchical information interaction between paths. The architecture of TaiChiNet is illustrated in Fig. 2, which comprises two U-Nets to unilaterally mine the patterns of lesions and backgrounds. Given a breast ultrasound image as the input I, the two U-Nets extract deep features of negative backgrounds F_n and positive lesions F_p, respectively, which can be defined as follows:

$$F_n = E_n(I; W_n^e), \tag{1}$$

$$F_p = E_p(I; W_p^e), \tag{2}$$

where E_n and E_p are the encoders of two U-Nets, and W_n^e and W_p^e are the encoding parameters, separately.

As the first phase of information interaction between the deep features F_n andF_p, a cross-attention module is designed to bilaterally enhance F_n and F_p by exploiting their complementary relationship. The cross-attention module is offered as:

$$(F_n^{'}, F_p^{'}) = CA(F_n, F_p), \tag{3}$$

where $F_n^{'}$ and $F_p^{'}$ denote the enhanced features associated with backgrounds and lesions, respectively. CA is the abbreviation for the cross-attention module. Then, $F_n^{'}$ and $F_p^{'}$ are fed into two independent decoders to predict the outputs $O_n^{'}$ and $O_p^{'}$, which are formulated as :

$$O_n^{'} = D_n(F_n^{'}; W_n^d), \tag{4}$$

$$O_p^{'} = D_p(F_p^{'}; W_p^d), \tag{5}$$

where W_n^d and W_p^d denote the parameters of decoders D_n and D_p, respectively.

As the second-phase information interaction between the shallow features $O_n^{'}$ and $O_p^{'}$, a complementary head including two MLPs [32]–[34] are applied to fine-tune the information for the two predictions $O_n^{'}$ and $O_p^{'}$.

The details of the sub-network U-Net, the cross-attention module, and the MLP are introduced in Sec. III-A, III-B, and III-C, respectively. As the third phase of information interaction, a complementary loss will be illustrated in Sec. IV.

A. Sub-network U-Net

To construct a simple yet efficient mutual-learning network, a lightweight sub-network U-Net [35] is adopted as the backbone of TaiChiNet. Due to the symmetrical encoder-decoder structure and skip connection, U-Net can effectively maintain spatial information and suppress the gradient vanishing. Fig. 3 illustrates the specifics of the sub-network U-Net. The kernel size of convolution is set to 3×3. The kernel size of the max-pooling operation is 2×2, and both the encoder and the decoder contain five layers. The convolutional layers have 32, 64, 128, 256, and 512 channels, respectively.

B. C^2-attention: Channel-to-spatial Cross Attention

In most fields of computer vision, convolution and attention operations have their own advantages and complement each other [36]–[38]. The attention mechanism can improve feature representation using visual grouping [39], [40], which arranges the symbols on a map to represent distinct classes of geographical characteristics and focuses on the patterns in one group while disregarding the patterns in the other groups. Fig. 4 gives the visualization of the visual grouping in the negative and positive paths, and shows that the symbols on a map are organized into different groups. From Fig. 4, we can see that: 1) according to the aforementioned analyses, the negative and positive paths would pay more attention to the normal backgrounds and the foreground lesions, respectively; 2) the two paths can both provide the grouped information to infer the lesion and background regions, separately; and 3) in the binary classification task of breast lesion segmentation, information redundancy of deep features exists between different channels including similar features in each path.

In light of the above observations, the information interaction between the two paths is feasible, and we argue that a cross-attention module should include the following characteristics: 1) due to the differences and commonality of the two paths given in the 1^{st} and 2^{nd} observations, channel fusion is important for converging their similar structured information; 2) following the channel fusion, spatial attention is essential for each path because of the differences shown in the 1^{st} observation; and 3) in each path, the feature channels should be compacted to reduce redundancy before feature fusion as suggested in the 3^{rd} observation.

Accordingly, we build a channel-to-spatial cross-attention module for segmenting breast lesions in ultrasound images. It is referred to C^2-attention and is different from the traditional cascaded channel and spatial attentions [41]. The primary

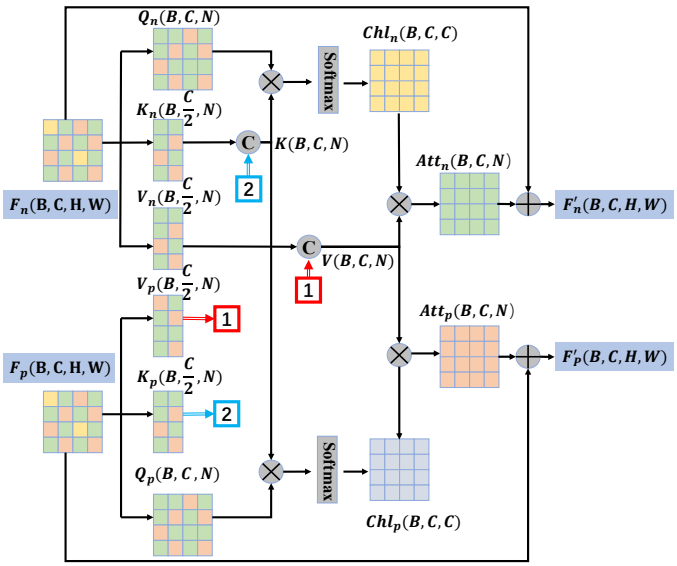

Fig. 5. Details of the channel-to-spatial cross-attention (C^2-attention) module, where B, C, H and W mean the batch size, number of channels, height, and width, respectively; and N=H×W. Two blocks '1' in red represent the connection nodes and the line between them is omitted for clarity. The two blue blocks '2' have the same representation.

explanation is that, in our C^2-attention, the channel fusion between the two paths serves the subsequent spatial attention rather than the original features. The architecture of C^2-attention module is given in Fig. 5, where F_n and F_p represent the deep features of the negative and positive U-Net paths of Fig. 2. After linear transformation, we obtain the query $Q_\triangle$, key $K_\triangle$ and value $V_\triangle$, where $\triangle \in \{n, p\}$. To reduce redundancy and meet the requirement of channel fusion, the channel numbers of key $K_\triangle$ and value $V_\triangle$ are reduced by half before the cross fusion.

In C^2-attention, K is obtained by concatenating K_n and K_p as (6). Then, we compute the dot products of Q_n and Q_p with K and apply the softmax function to generate the correlation scores Chl_n and Chl_p among channels in each U-Net path. By another dot product, the correlation scores Chl_n and Chl_p of channel fusion are weighted to value V, which is obtained by concatenating V_n and V_p as (7). Then, we get two spatial attention maps Att_n and Att_p for the negative and positive paths, respectively. Finally, the attention maps are added to the original feature F_n and F_p to enhance their representation ability. The C^2-attention module can be formulated as follow:

$$K = Concatenate(K_n, K_p), \tag{6}$$

$$V = Concatenate(V_n, V_p), \tag{7}$$

$$F_n^{'} = F_n + Softmax(Q_n K^T)V, \tag{8}$$

$$F_p^{'} = F_p + Softmax(Q_p K^T)V, \tag{9}$$

where $F_n^{'}$ and $F_p^{'}$ denote the features enhanced by C^2-attention.

C. Complementary Head

At the outputs of the two U-Nets, a complementary head is further designed to interact with each other. As a lightweight network, MLP is an option for the complementary head block. The multi-layer perceptron (MLP) is a simple but competitive network [43]–[45]. MLP receives a series of linearly projected feature maps as input, first raises its dimensionality initially and subsequently decreases it by linear layers.

In the complementary head shown as Fig. 2, two MLPs block is utilized to build a progressive refinement based on the complementary information between the negative and positive paths. The process can be defined as follows:

$$O_n = O_n^{'} + MLP(\Psi(O_p^{'}); W_n^m), \tag{10}$$

and

$$O_p = O_p^{'} + MLP(\Psi(O_n); W_p^m), \tag{11}$$

where O_n and O_p represent the final output of the negative and positive paths, respectively; W_n^m and W_p^m denote the parameters of two MLPs; Ψ means the complement operation as:

$$\Psi(A) = 1 - A, \tag{12}$$

where $A \in \{O_n, O_p^{'}\}$

IV. Objective Functions

To train TaiChiNet, we define three constraints on complementary supervision and lesion segmentation. Given the negative output O_n, the positive output O_p, the background ground truth G_n, and the lesion ground truth G_p, the objective function can be defined as

$$\mathbb{L} = L_S(O_n, G_n) + L_S(O_p, G_p) + L_C(O_n, O_p), \tag{13}$$

where the background ground truth G_n is the complementary set of G_p; $L_S(O_n, G_n)$ and $L_S(O_p, G_p)$ denote the segmentation constraints of backgrounds and lesions, respectively; and $L_C(O_n, O_p)$ represents the complementary loss. The complementary loss is introduced as the regularization term to enhance the generalization ability of the model. Next, we will introduce the constraints in detail.

A. Constraints on Segmentation

Following [5], we adopt a mixed loss that combines the dice coefficient loss l_d and binary cross-entropy loss l_b. The segmentation constraints can be defined as:

$$L_S(O_n, G_n) = l_d(O_n, G_n) + l_b(O_n, G_n), \tag{14}$$

$$L_S(O_p, G_p) = l_d(O_p, G_p) + l_b(O_p, G_p), \tag{15}$$

where

$$l_d(\widehat{\mathcal{Y}}, \mathcal{Y}) = 1 - \frac{2\sum\limits_{(i,j)} \mathcal{Y}(i,j)\widehat{\mathcal{Y}}(i,j)}{\sum\limits_{(i,j)} \mathcal{Y}(i,j)^2 + \sum\limits_{(i,j)} \widehat{\mathcal{Y}}(i,j)^2}, \tag{16}$$

$$l_b(\widehat{\mathcal{Y}}, \mathcal{Y}) = -\sum_{(i,j)} (1-\mathcal{Y}(i,j))log(1-\widehat{\mathcal{Y}}(i,j)) + \widehat{\mathcal{Y}}(i,j)log(\mathcal{Y}(i,j)), \tag{17}$$

where (i, j) represents the index of pixels; $\widehat{\mathcal{Y}} \in \{O_n, O_p\}$; and $\mathcal{Y} \in \{G_n, G_p\}$.

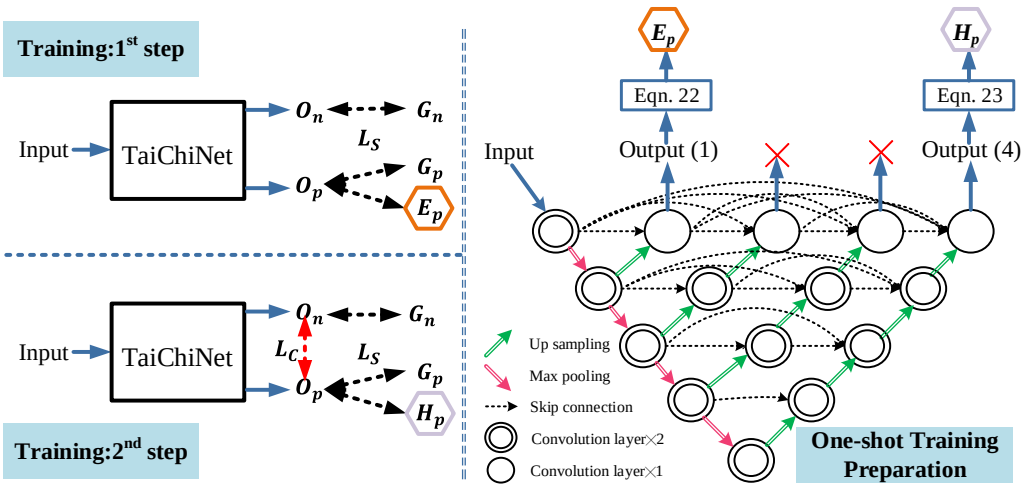

Fig. 6. Illustration of the training strategy of TaiChiNet. The one-shot training preparation based on U-Net++ [42] variant is utilized to provide $\boldsymbol{E_p}$ and $\boldsymbol{H_p}$. The One-shot preparation means that $\boldsymbol{E_p}$ and $\boldsymbol{H_p}$ are only mined once per dataset.

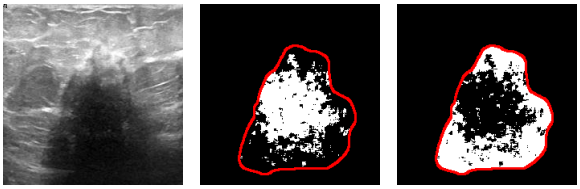

Fig. 7. Example of easy-to-hard strategy. (a) Ultrasound image. (b) Easy-part $\boldsymbol{E_P}$ of the lesion. (C) Hard-part $\boldsymbol{H_p}$ of the lesion. The red circle denotes the boundary of the lesion.

B. Constraints on Complementarity

The outputs of the background sub-network and lesion sub-network are complementary. It means that they should be consistent with each other after performing the complement operation as (12). The complementary loss $L_C(O_n, O_p)$ is defined based on the error in consistency, which is measured by a mean-square-error function (MSE). $L_C(O_n, O_p)$ is formulated as :

$$L_C(O_n, O_p) = \sum_{(i,j)} \frac{1}{2} MSE(O_n(i,j), 1 - O_p(i,j)) + \sum_{(i,j)} \frac{1}{2} MSE(O_p(i,j), 1 - O_n(i,j)), \tag{18}$$

where (i, j) represents the index of pixels, and MSE is defined as :

$$MSE(\hat{A}, A) = (\hat{A} - A)^2 \tag{19}$$

where A and $\hat{A} \in$ [0,1].

C. Training Strategy

Inspired by the progressive easy-to-hard learning [18], which splits targets into easy and hard portions, we design a progressive training technique to segment the breast lesions. The training strategy is shown in Fig. 6. Given the easy regions E_p and the hard regions H_p inside lesions as shown in Fig. 7, the training technique is also separated into two steps. E_p is used to warm the network at the first step, whereas the second one focuses on H_p. The weights are shared in the two steps. Therefore, (15) is reformulated as

$$L_S(O_p, G_p) = L_S^{1^{st}}(O_p, E_p, G_p) + L_S^{2^{nd}}(O_p, H_p, G_p), \tag{20}$$

where

$$L_S^{1^{st}}(O_p, E_p, G_p) = l_d(O_p, G_p) + l_b(O_p, G_p) + l_d(O_p, E_p) + l_b(O_p, E_p), \tag{21}$$

$$L_S^{2^{nd}}(O_p, H_p, G_p) = l_d(O_p, G_p) + l_b(O_p, G_p) + l_d(O_p, H_p) + l_b(O_p, H_p), \tag{22}$$

where the 1^{st} and 2^{nd} superscripts indicate the 1^{st} step and the 2^{nd} step, respectively. Additionally, $L_S(O_n, G_n)$ of (14) is used in both steps, but the complementary loss $L_C(O_n, O_p)$ only works in the 2^{nd} step.

To produce E_p and H_p, a model originated from U-Net++ [42] is applied since U-Net++ is integrated with different depths CNN and collaborative learning is embodied through aggregating multi-depth networks and supervising segmentation heads from each of the constituent networks. The outputs of U-Net++ may exhibit varying levels of segmentation, with coarse outputs at the first level and fine outputs at the fourth level. Therefore, the first-level and the fourth-level outputs are processed to discover E_p and H_p, respectively. The operation can be defined as :

$$E_p = Output(1) \cap GT, \tag{23}$$

$$H_p = (Output(4) \cap GT) \cap \overline{E_p}, \tag{24}$$

where $Output(1)$ and $Output(4)$ denote the first-level and fourth-level outputs of U-Net++, respectively. GT represents the ground truth of segmentation, and $\overline{E_p}$ is E_p's complementary set in GT. In the training strategy, the one-shot preparation refers to the generation of two parts: an easy part E_p and a hard part H_p using a U-Net++-based model. This

TABLE I
COMPARISON WITH DIFFERENT MODEL PARTS ON THE BUSI DATA SET. THE BEST VALUES ARE HIGHLIGHTED BY BOLD.

Models	Evolution	Params	DI	JA	ACC	PR	SE
Model 1	U-Net$_p$	8.64M	0.758 ± 0.027	0.668 ± 0.022	0.961 ± 0.006	0.771 ± 0.025	0.780 ± 0.023
Model 2	Model 1 + U-Net$_n$	17.28M	0.770 ± 0.012	0.678 ± 0.010	0.965 ± 0.003	0.783 ±0.014	0.801 ± 0.011
Model 3	Model 2 + CLoss	17.28M	0.774 ± 0.020	0.688 ± 0.015	0.966 ± 0.002	0.794 ± 0.016	0.813 ± 0.018
Model 4	Model 2 + CAtt	18.33M	0.791 ± 0.024	0.706 ± 0.014	0.975 ± 0.004	0.826 ± 0.020	0.831 ± 0.023
Model 5	Model 3 + CHead	17.28M	0.808 ± 0.029	0.715 ± 0.019	0.971 ± 0.004	0.819 ± 0.030	0.844 ± 0.023
Model 6	Model 3 + CAtt	18.33M	0.815 ± 0.023	0.724 ± 0.017	0.976 ± 0.005	0.817 ± 0.023	0.852 ± 0.024
Model 7	Model 5 + CHead	18.33M	0.823 ± 0.028	0.737 ± 0.016	0.979 ± 0.004	0.821 ± 0.022	0.865 ± 0.019
TaiChiNet	Model 5 + LS	18.33M	**0.836 ± 0.024**	**0.752 ± 0.023**	**0.986 ± 0.003**	**0.839 ± 0.027**	**0.878 ± 0.022**

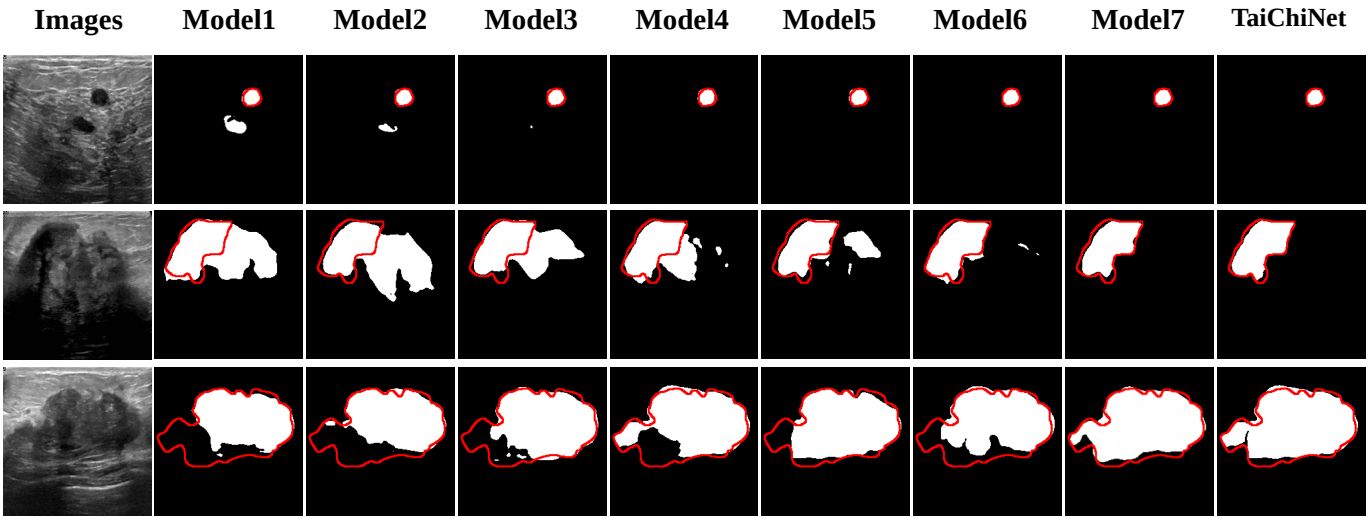

Fig. 8. Qualitative analyses among different versions of TaiChiNet. The models are introduced in Sec. V-D. The red circle denotes the boundary of ground truth.

process is performed only once per dataset, without the need for continuous updates or retraining. This approach reduces the complexity of the training session and makes it more practical for real-world applications.

V. EXPERIMENTS

To demonstrate the effectiveness of TaiChiNet, experiments are conducted to evaluate the performance of breast lesion segmentation. Two publicly available data sets (i.e., the BUSI data set [46] and the Dataset B [47]) are employed for training and testing. Sec. V-A presents introduction of the two datasets. The implementation details and evaluation metrics are presented in Sec. V-B and Sec. V-C, separately. The ablation analyses of model components are illustrated in Sec. V-D. In Sec. V-E and Sec. V-F, the mutual-learning ways and complementary heads are explored, respectively. Then, the comparisons with state-of-the-arts on the two data sets are described in Sec. V-G. In Sec. V-H, we investigate extended analyses in potential applications. Sec. V-I provides the qualitative visualization. Finally, the failure cases are discussed in Sec. V-J.

A. Datasets

The BUSI data set [46] contains 780 images collected from two types of ultrasound equipment in the Baheya Hospital, which includes 437 benign cases, 210 malignant cases, and 133 normal cases. The average image size of these images is 500 × 500 pixels. For fairness, the five-fold cross-validation is performed.

The Dataset B [47], which contains 110 images with benign lesions and 53 images with malignant lesions, is a public dataset collected from the UDIAT Diagnostic Center of the Parc Taulfi Corporation, Sabadell. The average image size is 760 × 570 pixels. The five-fold cross-validation is performed fairly.

B. Implementation Details

To increase the diversity of image samples, data augmentation is used including rotation (90°, 180°, 270°, and [−10°,10°]), gamma transformation (gamma ∈ [0.5,1.5]), and shear transformation (rate ∈ [0.6,1.3]). All input images are resized to 224 × 224, and the intensity of pixels is normalized to [0, 1]. All the experiments are implemented by Pytorch on an NVIDIA Geforce RTX 3090 GPU. In the training phase, an Adam optimizer with an initial learning rate 0.0001 is used to minimize the objective function. All networks are trained for 40 epochs, and the batch size is set to 16.

C. Evaluation Metric

To evaluate the segmentation capability of TaiChiNet, we introduce six widely used metrics, including Jaccard index (JA), Dice coefficient (DI), Accuracy (ACC), Sensitivity (SE),

TABLE II
COMPARISON BETWEEN TAICHINET WITH DIFFERENT MUTUAL LEARNING METHODS ON THE BUSI DATASET. CONS-ML AND COMP-ML REPRESENT CONSISTENT MUTUAL-LEARNING AND COMPLEMENTARY MUTUAL-LEARNING, RESPECTIVELY. THE BEST VALUES ARE HIGHLIGHTED BY BOLD.

Methods	DI	JA	ACC	PR	SE
U-Net$_p$	0.758 ± 0.027	0.668 ± 0.022	0.961 ± 0.006	0.771 ± 0.025	0.780 ± 0.023
TaiChiNet w Cons-ML	0.813 ± 0.029	0.731 ± 0.027	0.978 ± 0.005	0.817 ± 0.031	0.843 ± 0.025
TaiChiNet w Comp-ML	**0.836 ± 0.024**	**0.752 ± 0.023**	**0.986 ± 0.003**	**0.839 ± 0.027**	**0.878 ± 0.022**

and Precision (PR). They are defined as follows:

$$JA = \frac{TP}{TP + FP + FN}, \tag{25}$$

$$DI = \frac{2 \times TP}{2 \times TP + FP + FN}, \tag{26}$$

$$ACC = \frac{TP + TN}{TP + FP + FN + TN}, \tag{27}$$

$$SE = \frac{TP}{TP + FN}, \tag{28}$$

$$PR = \frac{TP}{TP + FP} \tag{29}$$

where TP (True Positive) and TN (True Negative) represent the number of foreground pixels and background pixels correctly segmented; FP (False Positive) denotes the background pixels that are incorrectly labeled as the foreground pixels; FN (False Negative) denotes the foreground pixels that are incorrectly predicted as the background pixels.

D. Ablation Analyses of Model Components

To demonstrate the effectiveness of different model components in TaiChiNet, we conduct an ablation study on the BUSI data set. In summary, the following models are compared:

Model 1: the positive lesion path with a single U-Net, which is denoted as U-Net$_p$.

Model 2: on the basis of Model 1, the negative path U-Net$_n$ is added without interaction, and the segmentation results are calculated by the average value of the outputs of the two paths.

Model 3: the complementary loss (+CLoss) is added in Model 2.

Model 4: the cross-attention module (+CAtt) is added in Model 2.

Model 5: the complementary head (+CHead) with two MLPs is introduced into Model 3.

Model 6: the cross-attention module (+CAtt) is employed in Model 3.

Model 7: the complementary head (+CHead) with two MLPs is introduced into Model 5.

TaiChiNet: the designed easy-to-hard learning-strategy (+LS) is used in Model 7.

The results are listed in Table I. It can be seen that:

• TaiChiNet beats the basic models Model 1 and Model 2 by a large margin. This indicates that introducing complementary information between the foreground lesions and normal backgrounds can effectively improve performance.

• Because of the complementary loss, Model 3 outperforms Model 2. The main reason is that the complementary loss would act as a regularization term to enhance the model's generalization ability.

• Due to the cross-attention module, the performance of Model 4 is substantially enhanced. It implies that the cross-attention module can assist the model with information fusion between the negative and positive paths in the deep features.

• Thanks to the complementary head using two MLPs, Model 5 outperforms Model 4 on the BUSI data set. It demonstrates that the complementary information between the shallow features is also useful for lesion segmentation.

• Benefiting from the easy-to-hard learning strategy, TaiChiNet further achieves an improvement.

Fig. 8 presents the visualization comparison of TaiChiNet with different components. We can observe that each component of TaiChiNet is beneficial to performance improvement.

Fig. 9 gives the comparison of TaiChiNet without and with the cross-attention module. It shows that the cross-attention module can enhance the feature compactness of foreground lesions and restrains the response of normal backgrounds.

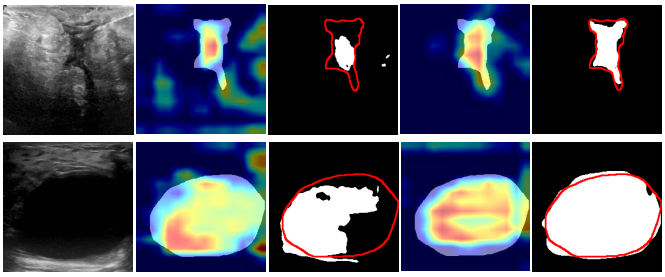

Fig. 9. Qualitative analyses of TaiChiNet without and with cross-attention module. (a) Ultrasound images. (b) Feature maps of TaiChiNet without cross-attention module. (c) Outputs of TaiChiNet without cross-attention module. (d) Feature maps enhanced by cross-attention module. (e) Outputs of TaiChiNet with cross-attention module. The red circle denotes the boundary of ground truth.

E. Comparative Analyses of Different Mutual-learning Methods

In this section, we give two kinds of mutual-learning methods. One is consistent mutual learning between two paths, which both focus on the foreground lesions. It is used to demonstrate the effect of simply increasing parameters. The other one is TaiChiNet with the complementary mutual learning between the foreground lesions and the normal backgrounds. The experimental results are listed in Table II. It can be seen that:

• TaiChiNet with two kinds of mutual-learning methods both can improve the performance of the lesion segmentation.

TABLE III
COMPARISON BETWEEN TAICHINET WITH DIFFERENT COMPLEMENTARY HEADS ON THE BUSI DATA SET. THE BEST VALUES ARE HIGHLIGHTED BY BOLD.

Methods	DI	JA	ACC	PR	SE
TaiChiNet with (a)	0.821 ± 0.027	0.736 ± 0.019	0.974 ± 0.005	0.844 ± 0.024	0.836 ± 0.025
TaiChiNet with (b)	0.827 ± 0.025	0.743 ± 0.022	0.978 ± 0.006	**0.849 ± 0.026**	0.854 ± 0.023
TaiChiNet with (c)	**0.836 ± 0.024**	**0.752 ± 0.023**	**0.986 ± 0.003**	0.839 ± 0.027	**0.878 ± 0.022**

* TaiChiNet with ($\triangle$) represents TaiChiNet with the head provided in Fig. 10($\triangle$), $\triangle \in \{a, b, c\}$.

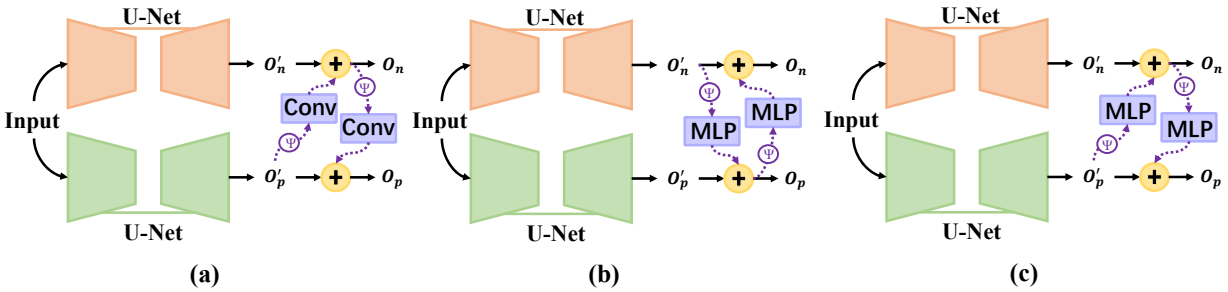

Fig. 10. Examples of different complementary heads: (a) TaiChiNet with Conv-head ($O_p^{'} \rightarrow O_n \rightarrow O_p$); (b) TaiChiNet with MLP-head ($O_n^{'} \rightarrow O_p \rightarrow O_n$); (c) TaiChiNet with MLP-head ($O_p^{'} \rightarrow O_n \rightarrow O_p$).

TABLE IV
COMPARISONS WITH STATE-OF-THE-ART METHODS FOR BREAST LESION SEGMENTATION ON BUSI DATASET. THE FIRST PART CONTAINS THE RESULTS WITHOUT NORMAL CASES, WHEREAS THE SECOND PART INCLUDES NORMAL CASES. THE BEST VALUES ARE HIGHLIGHTED BY BOLD.

Methods	DI	JA	ACC	PR	SE
U-Net [35]	0.725 ± 0.032	0.645 ± 0.031	0.971± 0.004	0.723 ± 0.033	0.779 ± 0.032
U-Net++ [42]	0.737 ± 0.031	0.656 ± 0.030	0.973 ± 0.004	0.735 ± 0.032	0.781 ± 0.032
attention U-Net [48]	0.755 ± 0.030	0.624 ± 0.029	0.973 ± 0.004	0.752 ± 0.031	0.796 ± 0.030
deeplabv3 [49]	0.777 ± 0.027	0.691 ± 0.026	0.974 ± 0.004	0.782 ± 0.027	0.806 ± 0.026
CE-Net [50]	0.753 ± 0.028	0.665 ± 0.028	0.974 ± 0.004	0.767 ± 0.029	0.786 ± 0.02
FAT-Net [51]	0.784 ± 0.025	0.694 ± 0.025	0.974 ± 0.004	0.782 ± 0.025	0.819 ± 0.026
X-Net [52]	0.810 ± 0.026	0.728± 0.024	0.976 ± 0.003	0.808 ± 0.024	0.813 ± 0.020
Wang *et al.* [53]	0.771 ± 0.028	0.684 ± 0.028	0.973 ± 0.004	0.804 ± 0.030	0.807 ± 0.027
DE-ResUnet [54]	0.763 ± 0.034	0.687 ± 0.019	0.972 ± 0.006	0.771 ± 0.034	0.783 ± 0.025
AAU-Net [55]	0.781 ± 0.025	0.704 ± 0.026	0.974 ± 0.004	0.795 ± 0.023	0.812 ± 0.027
TaiChiNet	**0.836 ± 0.024**	**0.752 ± 0.023**	**0.986 ± 0.003**	**0.839 ± 0.027**	**0.878 ± 0.022**
U-Net [35]	0.587 ± 0.016	0.529 ± 0.021	0.959 ± 0.002	0.581 ± 0.012	0.636 ± 0.011
U-Net++ [42]	0.598 ± 0.021	0.545 ± 0.018	0.962 ± 0.003	0.596 ± 0.017	0.641 ± 0.019
attention U-Net [48]	0.603 ± 0.023	0.556 ± 0.037	0.961 ± 0.007	0.612 ± 0.026	0.665 ± 0.034
deeplabv3 [49]	0.623 ± 0.031	0.568 ± 0.025	0.963 ± 0.002	0.627 ± 0.019	0.683 ± 0.024
CE-Net [50]	0.628 ± 0.026	0.576 ± 0.028	0.962 ± 0.004	0.623 ± 0.023	0.667 ± 0.038
FAT-Net [51]	0.631 ± 0.028	0.589 ± 0.031	0.964 ± 0.005	0.635 ± 0.025	0.687 ± 0.032
X-Net [52]	0.651 ± 0.024	0.588 ± 0.025	0.966 ± 0.005	0.655 ± 0.028	0.717 ± 0.029
Wang *et al.* [53]	0.617 ± 0.026	0.558 ± 0.024	0.964 ± 0.006	0.633 ± 0.024	0.682 ± 0.014
DE-ResUnet [54]	0.619 ± 0.027	0.567 ± 0.027	0.962 ± 0.005	0.626 ± 0.026	0.679 ± 0.025
AAU-Net [55]	0.624 ± 0.024	0.562 ± 0.022	0.964 ± 0.002	0.629 ± 0.021	0.685 ± 0.016
TaiChiNet	**0.673 ± 0.027**	**0.624 ± 0.029**	**0.969 ± 0.004**	**0.675 ± 0.026**	**0.738 ± 0.028**

It infers that information interaction is important.

• Specifically, TaiChiNet with complementary information has the highest performance, confirming once again that complementary information from backgrounds is advantageous for reducing misunderstanding.

F. Comparative Analyses of Different Complementary Heads

In this section, the comparative analyses of different complementary heads are conducted. Convolutions and MLPs are investigated in Fig. 10, and the results are listed in Table III. It can be observed that TaiChiNet with MLP-head ($O_p^{'} \rightarrow O_n \rightarrow O_p$) generally outperforms the others. Thus, the third architecture of the complementary head is adopted in our TaiChiNet.

G. Comparison with State-of-the-arts

In this section, we compare TaiChiNet against several deep-learning-based segmentation methods, including U-Net [35], U-Net++ [42], attention U-Net [48], deeplabv3 [49], CE-Net [50], FAT-Net [51], X-Net [52], Wang *et al.* [53], DE-ResUnet [54] and AAU-Net [55]. To provide fair comparisons, we obtain the segmentation results of the compared methods with the same implementation details and their networks are retrained on the datasets.

Quantitative comparisons on BUSI. Table IV illustrates the mean and standard deviation values for TaiChiNet and all comparison methods on the BUSI dataset. Considering that normal and lesion ultrasound images are both processed in clinical breast ultrasound analysis, we conduct two experiments with or without the normal cases of the BUSI dataset in

TABLE V
COMPARISONS WITH STATE-OF-THE-ART METHODS FOR BREAST LESION SEGMENTATION ON DATASET B. THE FIRST PART LISTS THE RESULTS WITHIN DATASET B. IN THE SECOND PART, THE MODEL IS TRAINED ON THE BUSI DATASET AND EVALUATED ON DATASET B. THE BEST VALUES ARE HIGHLIGHTED BY BOLD.

Methods	DI	JA	ACC	PR	SE
U-Net [35]	0.785 ± 0.027	0.702 ± 0.027	0.975 ±0.002	0.772 ± 0.028	0.835 ± 0.028
U-Net++ [42]	0.789 ± 0.026	0.713 ± 0.025	0.978 ± 0.002	0.788 ± 0.027	0.841 ± 0.026
attention U-Net [48]	0.791 ± 0.024	0.725 ± 0.024	0.984 ± 0.003	0.806 ± 0.023	0.849 ± 0.026
deeplabv3 [49]	0.798 ± 0.018	0.721 ± 0.020	0.989 ± 0.001	0.811 ± 0.020	0.850 ± 0.020
CE-Net [50]	0.813 ± 0.019	0.717 ± 0.021	0.989 ± 0.002	0.821 ± 0.021	0.845 ± 0.022
FAT-Net [51]	0.801 ± 0.024	0.720 ± 0.024	0.984 ± 0.003	0.808 ± 0.024	0.869 ± 0.022
X-Net [52]	0.831 ± 0.017	0.755± 0.024	0.986 ± 0.013	0.834 ± 0.022	0.863 ± 0.025
Wang *et al.* [53]	0.820 ± 0.021	0.750 ± 0.022	0.986 ± 0.014	0.824 ± 0.025	0.862 ± 0.021
DE-ResUnet [54]	0.796 ± 0.029	0.720 ± 0.025	0.987 ± 0.013	0.808 ± 0.024	0.847 ± 0.024
AAU-Net [55]	**0.842 ± 0.018**	0.759 ± 0.019	0.989 ± 0.002	0.845 ± 0.019	0.865 ± 0.020
TaiChiNet	0.842 ± 0.024	**0.775 ± 0.024**	**0.990 ± 0.002**	**0.856 ± 0.023**	**0.912 ± 0.001**
U-Net [35]	0.715 ± 0.028	0.644 ± 0.027	0.979 ± 0.002	0.714 ± 0.031	0.779 ± 0.027
U-Net++ [42]	0.723 ± 0.027	0.647 ± 0.027	0.983 ± 0.002	0.749 ± 0.032	0.784 ± 0.028
attention U-Net [48]	0.734 ± 0.030	0.646 ± 0.028	0.985 ± 0.002	0.722 ± 0.032	0.822 ± 0.028
deeplabv3 [49]	0.737 ± 0.029	0.647 ± 0.028	0.983 ± 0.002	0.744 ± 0.027	0.787 ± 0.029
CE-Net [50]	0.732 ± 0.032	0.643 ± 0.031	0.979 ± 0.002	0.759 ± 0.034	0.788 ± 0.033
FAT-Net [51]	0.748 ± 0.028	0.660 ± 0.028	0.984 ± 0.002	0.755 ± 0.028	0.783 ± 0.029
X-Net [52]	0.776 ± 0.028	0.673 ± 0.026	0.981 ± 0.004	0.765 ± 0.027	0.796 ± 0.023
Wang *et al.* [53]	0.782 ± 0.031	0.664 ± 0.030	0.981 ± 0.002	0.771 ± 0.034	0.801 ± 0.029
DE-ResUnet [54]	0.735 ± 0.024	0.646 ± 0.026	0.981 ± 0.005	0.757 ± 0.028	0.785 ± 0.021
AAU-Net [55]	0.781 ± 0.026	0.670 ± 0.021	0.983 ± 0.004	0.775 ± 0.022	0.796 ± 0.012
TaiChiNet	**0.792 ± 0.024**	**0.696 ± 0.025**	**0.985 ± 0.002**	**0.787 ± 0.002**	**0.840 ± 0.023**

TABLE VI
COMPARISONS WITH STATE-OF-THE-ART METHODS FOR THYROID NODULE SEGMENTATION ON TN-SCUI2020. THE BEST VALUES ARE HIGHLIGHTED BY BOLD.

Methods	DI	JA	ACC	PR	SE
U-Net [35]	0.835 ± 0.018	0.754 ± 0.025	0.972 ± 0.004	0.830 ± 0.020	0.869 ± 0.019
U-Net++ [42]	0.838 ± 0.020	0.765 ± 0.021	0.973 ± 0.003	0.833 ± 0.018	0.873 ± 0.019
attention U-Net [48]	0.847 ± 0.018	0.775 ± 0.004	0.974 ± 0.002	0.845 ± 0.020	0.880 ± 0.015
deeplabv3 [49]	0.854 ± 0.013	0.787 ± 0.019	0.976 ± 0.003	0.854 ± 0.019	0.881 ± 0.017
CE-Net [50]	0.850 ± 0.018	0.785 ± 0.020	0.976 ± 0.002	0.861 ± 0.017	0.877 ± 0.017
FAT-Net [51]	0.854 ± 0.016	0.795 ± 0.019	0.979 ± 0.002	0.863 ± 0.017	0.880 ± 0.016
Wang *et al.* [53]	0.849 ± 0.017	0.763 ± 0.020	0.975 ± 0.003	0.842 ± 0.019	0.881 ± 0.019
AAU-Net [55]	0.845 ± 0.018	0.761 ± 0.020	0.975 ± 0.004	0.848 ± 0.018	0.879 ± 0.018
TaiChiNet	**0.885 ± 0.019**	**0.805 ± 0.002**	**0.984 ± 0.003**	**0.881 ± 0.013**	**0.908 ± 0.019**

the training and test phase, respectively. The first part contains the results without normal cases, whereas the second part includes normal cases. It can be observed as:

- Compared with other segmentation methods on BUSI without normal cases, the performance of our proposed TaiChiNet improves on all five metrics. It demonstrates that TaiChiNet including complementary information can discriminate breast lesions from backgrounds more accurately than all competitors.
- When normal cases are included, TaiChiNet has the highest segmentation performance. The main reason would be that the normal-background path limits false positives to some degree.

Quantitative comparisons on Dataset B. The first part of Table V lists the comparison results on Dataset B. In the second part of Table V, cross-validation between BUSI and Dataset B, i.e., the model trained on BUSI and evaluated on Dataset B, is used to evaluate the generalization capacity further. We can see that TaiChiNet still achieves top performance. Due to the mutual-learning mechanism, TaiChiNet obtains a competitive performance of domain adaptation.

H. Extended Analyses in Potential Applications

To illustrate the potential value of TaiChiNet, we apply it on the ultrasound thyroid nodule segmentation of the TN-SCUI2020 dataset [3]and retinal optic cup/disc segmentation of the REFUGE dataset [4]. Considering that the ground truth of the test set on the TN-SCUI2020 dataset is not available, we randomly selected 20% of the training data as the test set. It is subjected to five-fold cross-validation to ensure fairness. The REFUGE dataset provides predefined partitions with 400, 400, and 400 for training, validation, and testing, respectively.

Comparison results on TN-SCUI2020. Table VI lists the comparison results between TaiChiNet and the competitors. It can be seen that TaiChiNet obtains the best segmentation performance, which demonstrates that TaiChiNet can distinguish thyroid nodules from ultrasound pictures efficiently.

Comparison results on REFUGE. Table VII presents the comparison results of the optic disc and cup segmentation. Again, compared to other segmentation methods, TaiChiNet delivers performance benefits.

[3]https://tn-scui2020.grand-challenge.org/
[4]https://refuge.grand-challenge.org/.

TABLE VII
COMPARISONS WITH STATE-OF-THE-ART METHODS FOR RETINAL OPTIC CUP AND DISC SEGMENTATION ON REFUGE. THE BEST VALUES ARE HIGHLIGHTED BY BOLD.

Methods	DI (disc)	JA (disc)	DI (cup)	JA (cup)
U-Net [35]	0.939 ± 0.002	0.906 ± 0.003	0.876 ± 0.003	0.781 ± 0.004
U-Net++ [42]	0.942 ± 0.008	0.910 ± 0.007	0.878 ± 0.006	0.790 ± 0.005
attention U-Net [48]	0.945 ± 0.005	0.918 ± 0.006	0.862 ± 0.002	0.798 ± 0.006
deeplabv3 [49]	0.952 ± 0.002	0.926 ± 0.009	0.872 ± 0.005	0.796 ± 0.008
CE-Net [50]	0.946 ± 0.006	0.917 ± 0.007	0.861 ± 0.006	0.793 ± 0.009
FAT-Net [51]	0.951 ± 0.004	0.919 ± 0.006	0.868 ± 0.005	0.798 ± 0.008
Wang *et al.* [53]	0.957 ± 0.013	0.918 ± 0.012	0.859 ± 0.018	0.801 ± 0.015
AAU-Net [55]	0.962 ± 0.019	0.927 ± 0.034	0.887 ± 0.034	0.802 ± 0.019
TaiChiNet	**0.974 ± 0.001**	**0.941 ± 0.003**	**0.898 ± 0.006**	**0.839 ± 0.007**

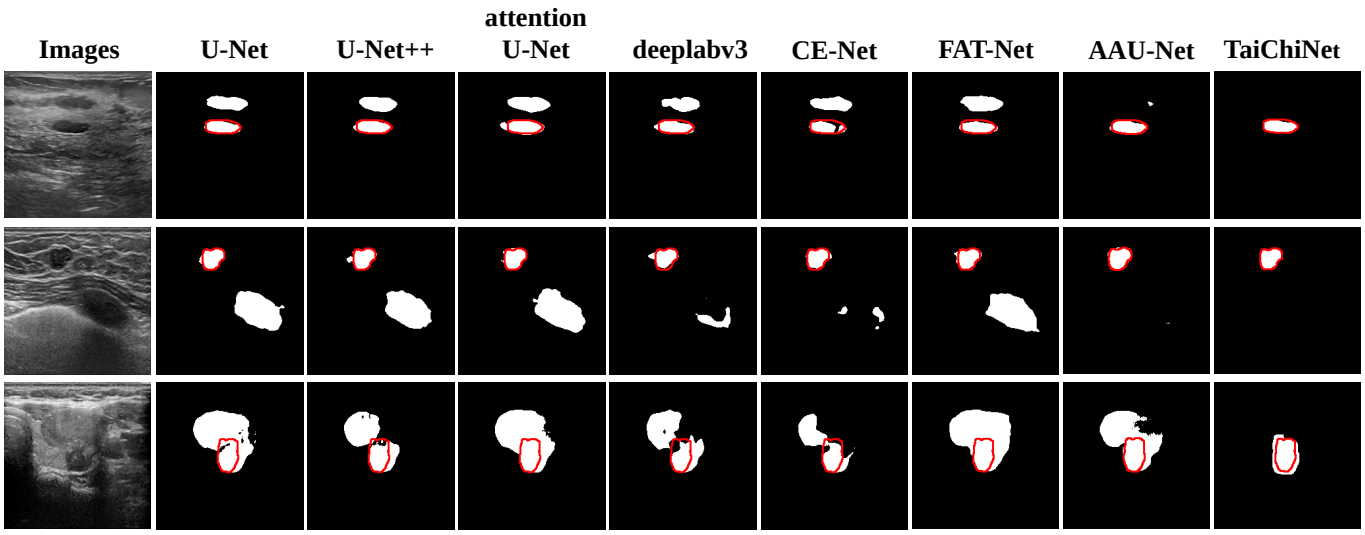

Fig. 11. Qualitative visualization. The first line is the segmentation results on BUSI dataset. The second line is the segmentation results on Dataset B. The third line is the segmentation results on TN-SCUI2020. The red circle denotes the boundary of ground truth.

I. Qualitative Visualization

The qualitative visualization on the BUSI, Dataset B, and TN-SCUI2020 datasets is illustrated in Fig. 11. It can be seen that TaiChiNet outperforms all competitors. Compared to FAT-Net, which employs both a transformer encoder and CNN encoder, TaiChiNet can distinguish between the foreground and background more precisely and generates more accurate segmentation results. The main reason is that TaiChiNet employs a cross-attention mechanism, which could aid in preventing TaiChiNet from being confused by background distractions.

J. Discussion about failure cases

We will discuss two scenarios in which the performance of TaiChiNet will be weakened: 1) in the first row of Fig. 12, large lesions with very extremely similar appearances to normal backgrounds; 2) in the second row of Fig. 12, large lesions with contradictory information compared to typical samples. The former situation would lead to false negatives. In the latter one, the lesions and backgrounds are about proportional, and the black regions are normal backgrounds. In several samples, lesions often take the attribute of black regions. Therefore, TaiChiNet provides an opposite prediction outcome in the latter situation. In the future, we will introduce sample-level curriculum learning to deal with these hard samples. The objective of the sample-level learning curriculum is to train a deep learning model with data of progressively increasing complexity. Since the model has already been trained using basic examples, there are greater opportunities to improve its performance when presented with complex data [18].

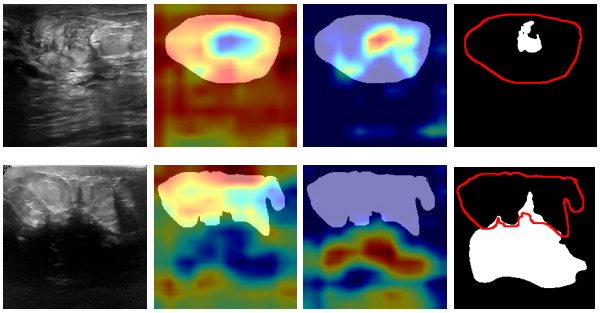

Fig. 12. Failure cases. (a) Ultrasound images. (b) Features in the negative path. (c) Features in the positive path. (d) The predicted outputs. The red circle denotes the boundary of ground truth, and the regions of ground truth are highlighted.

VI. CONCLUSION

In an effort to improve breast lesion segmentation by reducing false positives in the normal background and false

negatives in the foreground lesions, we propose a negative-positive cross-attention network to simultaneously enhance the segmentation performance of the normal background and the foreground lesions. The network is designated as TaiChiNet, which is derived from the complementing opposites of bipolarity in TaiChi. TaiChiNet models the task of breast lesion segmentation as a mutual-learning task via multi-level information interactions at the deep-feature layer, the shallow-feature layer, and the supervision. For the multi-level information interactions, a channel-to-spatial cross-attention, a complimentary head, and a complementary loss are designed accordingly. In addition, taking into account the varying segmentation difficulties of pixels within foreground lesions, we introduce a pixel-level easy-to-hard learning strategy to assist model learning in the training phase for further performance gains. Experimental results demonstrate the effectiveness of TaiChiNet in breast lesion segmentation. Future research will focus on reducing false positives and false negatives in a variety of prospective applications.

REFERENCES

[1] A. Ahmad, *Breast cancer metastasis and drug resistance: challenges and progress*. Springer, 2019, vol. 1152.

[2] W. A. Berg, J. D. Blume, J. B. Cormack, E. B. Mendelson, D. Lehrer, M. Böhm-Vélez, E. D. Pisano, R. A. Jong, W. P. Evans, M. J. Morton *et al.*, "Combined screening with ultrasound and mammography vs mammography alone in women at elevated risk of breast cancer," *Jama*, vol. 299, no. 18, pp. 2151–2163, 2008.

[3] W. A. Berg, Z. Zhang, D. Lehrer, R. A. Jong, E. D. Pisano, R. G. Barr, M. Böhm-Vélez, M. C. Mahoney, W. P. Evans, L. H. Larsen *et al.*, "Detection of breast cancer with addition of annual screening ultrasound or a single screening mri to mammography in women with elevated breast cancer risk," *Jama*, vol. 307, no. 13, pp. 1394–1404, 2012.

[4] Y. Xu, Y. Wang, J. Yuan, Q. Cheng, X. Wang, and P. L. Carson, "Medical breast ultrasound image segmentation by machine learning," *Ultrasonics*, vol. 91, pp. 1–9, 2019.

[5] C. Xue, L. Zhu, H. Fu, X. Hu, X. Li, H. Zhang, and P.-A. Heng, "Global guidance network for breast lesion segmentation in ultrasound images," *Medical image analysis*, vol. 70, p. 101989, 2021.

[6] G. Chen, Y. Dai, J. Zhang, and M. H. Yap, "Aau-net: An adaptive attention u-net for breast lesions segmentation in ultrasound images," *arXiv preprint arXiv:2204.12077*, 2022.

[7] A. Iqbal and M. Sharif, "Mda-net: Multiscale dual attention-based network for breast lesion segmentation using ultrasound images," *Journal of King Saud University-Computer and Information Sciences*, 2021.

[8] Y. Hu, Y. Guo, Y. Wang, J. Yu, J. Li, S. Zhou, and C. Chang, "Automatic tumor segmentation in breast ultrasound images using a dilated fully convolutional network combined with an active contour model," *Medical physics*, vol. 46, no. 1, pp. 215–228, 2019.

[9] J. Li, L. Cheng, T. Xia, H. Ni, and J. Li, "Multi-scale fusion u-net for the segmentation of breast lesions," *IEEE Access*, vol. 9, pp. 137 125–137 139, 2021.

[10] R. Irfan, A. A. Almazroi, H. T. Rauf, R. Damaševičius, E. A. Nasr, and A. E. Abdelgawad, "Dilated semantic segmentation for breast ultrasonic lesion detection using parallel feature fusion," *Diagnostics*, vol. 11, no. 7, p. 1212, 2021.

[11] N. Abraham and N. M. Khan, "A novel focal tversky loss function with improved attention u-net for lesion segmentation," in *2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019)*. IEEE, 2019, pp. 683–687.

[12] H. Lee, J. Park, and J. Y. Hwang, "Channel attention module with multiscale grid average pooling for breast cancer segmentation in an ultrasound image," *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, vol. 67, no. 7, pp. 1344–1353, 2020.

[13] M. Yeung, E. Sala, C.-B. Schönlieb, and L. Rundo, "Unified focal loss: Generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation," *Computerized Medical Imaging and Graphics*, vol. 95, p. 102026, 2022.

[14] A. Sinha and J. Dolz, "Multi-scale self-guided attention for medical image segmentation," *IEEE journal of biomedical and health informatics*, vol. 25, no. 1, pp. 121–130, 2020.

[15] Q. Yan, B. Wang, W. Zhang, C. Luo, W. Xu, Z. Xu, Y. Zhang, Q. Shi, L. Zhang, and Z. You, "Attention-guided deep neural network with multi-scale feature fusion for liver vessel segmentation," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 7, pp. 2629–2642, 2020.

[16] K. Chen, L. Yao, D. Zhang, X. Chang, G. Long, and S. Wang, "Distributionally robust semi-supervised learning for people-centric sensing," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 3321–3328.

[17] C.-Y. Hsu, M. O'Connor, and S. Lee, "Understandings of death and dying for people of chinese origin," *Death studies*, vol. 33, no. 2, pp. 153–174, 2009.

[18] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proceedings of the 26th annual international conference on machine learning*, 2009, pp. 41–48.

[19] P. Tang, X. Yan, Q. Liang, and D. Zhang, "Afln-dgcl: Adaptive feature learning network with difficulty-guided curriculum learning for skin lesion segmentation," *Applied Soft Computing*, vol. 110, p. 107656, 2021.

[20] F. Liu, S. Ge, and X. Wu, "Competence-based multimodal curriculum learning for medical report generation," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 3001–3012.

[21] X. Dong, C. Long, W. Xu, and C. Xiao, "Dual graph convolutional networks with transformer and curriculum learning for image captioning," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 2615–2624.

[22] Y. Liu, J. Zhou, L. Liu, Z. Zhan, Y. Hu, Y. Fu, and H. Duan, "Fcp-net: a feature-compression-pyramid network guided by game-theoretic interactions for medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 41, no. 6, pp. 1482–1496, 2022.

[23] B. Shareef, A. Vakanski, M. Xian, and P. E. Freer, "Estan: Enhanced small tumor-aware network for breast ultrasound image segmentation," *arXiv preprint arXiv:2009.12894*, 2020.

[24] Z. Yang, Y. Wei, and Y. Yang, "Collaborative video object segmentation by foreground-background integration," in *European Conference on Computer Vision*. Springer, 2020, pp. 332–348.

[25] Y. Oh, B. Kim, and B. Ham, "Background-aware pooling and noise-aware loss for weakly-supervised semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6913–6922.

[26] Y. Yang, H. Bilen, Q. Zou, W. Y. Cheung, and X. Ji, "Learning foreground-background segmentation from improved layered gans," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 2524–2533.

[27] B. Sauvalle and A. de La Fortelle, "Autoencoder-based background reconstruction and foreground segmentation with background noise estimation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 3244–3255.

[28] L. Wu, S. Hu, and C. Liu, "Mr brain segmentation based on de-resunet combining texture features and background knowledge," *Biomedical Signal Processing and Control*, vol. 75, p. 103541, 2022.

[29] 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, 2021.

[30] L. Zhu, R. Chen, H. Fu, C. Xie, L. Wang, L. Wan, and P.-A. Heng, "A second-order subregion pooling network for breast lesion segmentation in ultrasound," pp. 160–170, 2020.

[31] N. Abraham and N. M. Khan, "A novel focal tversky loss function with improved attention u-net for lesion segmentation," in *2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019)*. IEEE, 2019, pp. 683–687.

[32] Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Bovik, and Y. Li, "Maxim: Multi-axis mlp for image processing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5769–5780.

[33] T. Yu, X. Li, Y. Cai, M. Sun, and P. Li, "S2-mlp: Spatial-shift mlp architecture for vision," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 297–306.

[34] S. Chen, E. Xie, C. Ge, D. Liang, and P. Luo, "Cyclemlp: A mlp-like architecture for dense prediction," *arXiv preprint arXiv:2107.10224*, 2021.

[35] 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.
[36] Q. Han, Z. Fan, Q. Dai, L. Sun, M.-M. Cheng, J. Liu, and J. Wang, "On the connection between local attention and dynamic depth-wise convolution," in *International Conference on Learning Representations*, 2021.
[37] N. Park and S. Kim, "How do vision transformers work?" *arXiv preprint arXiv:2202.06709*, 2022.
[38] Q. Yu, Y. Xia, Y. Bai, Y. Lu, A. L. Yuille, and W. Shen, "Glance-and-gaze vision transformer," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12 992–13 003, 2021.
[39] D. Zhou, Z. Yu, E. Xie, C. Xiao, A. Anandkumar, J. Feng, and J. M. Alvarez, "Understanding the robustness in vision transformers," *arXiv preprint arXiv:2204.12451*, 2022.
[40] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9650–9660.
[41] 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 (CVPR)*, June 2019.
[42] 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, 2020.
[43] H. Liu, Z. Dai, D. So, and Q. V. Le, "Pay attention to mlps," *Advances in Neural Information Processing Systems*, vol. 34, pp. 9204–9215, 2021.
[44] I. O. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit *et al.*, "Mlp-mixer: An all-mlp architecture for vision," *Advances in Neural Information Processing Systems*, vol. 34, pp. 24 261–24 272, 2021.
[45] S. Chen, E. Xie, C. Ge, D. Liang, and P. Luo, "Cyclemlp: A mlp-like architecture for dense prediction," *arXiv preprint arXiv:2107.10224*, 2021.
[46] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, "Dataset of breast ultrasound images," *Data in brief*, vol. 28, p. 104863, 2020.
[47] 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*, vol. 107, p. 101880, 2020.
[48] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz *et al.*, "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.
[49] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.
[50] 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*, vol. 38, no. 10, pp. 2281–2292, 2019.
[51] H. Wu, S. Chen, G. Chen, W. Wang, B. Lei, and Z. Wen, "Fat-net: Feature adaptive transformers for automated skin lesion segmentation," *Medical Image Analysis*, vol. 76, p. 102327, 2022.
[52] Y. Li, Z. Wang, L. Yin, Z. Zhu, G. Qi, and Y. Liu, "X-net: a dual encoding–decoding method in medical image segmentation," *The Visual Computer*, pp. 1–11, 2021.
[53] K. Wang, S. Liang, and Y. Zhang, "Residual feedback network for breast lesion segmentation in ultrasound image," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*. Springer, 2021, pp. 471–481.
[54] L. Wu, S. Hu, and C. Liu, "Mr brain segmentation based on de-resunet combining texture features and background knowledge," *Biomedical Signal Processing and Control*, vol. 75, p. 103541, 2022.
[55] 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*, 2022.

Jinting Wang received her M.S. degree from the School of Biomedical Engineering at Southern Medical University, Guangzhou, China, in 2023, where she had also earned a BS degree in 2020. She is currently pursuing a PhD degree within the Artificial Intelligence Thrust at the Information Hub, Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China. Her current research interests include multimodal learning, Artificial Intelligence Generated Content (AIGC), medical image processing, and deep learning.

Jiafei Liang received her M.S. degree from the School of Biomedical Engineering at Southern Medical University, Guangzhou, China, in 2023, where she had also earned a BS degree in 2019. Her research interests include anomaly detection and deep learning.

Yang Xiao received the B.S., M.S., and Ph.D. degrees from the Huazhong University of Science and Technology, Wuhan, China, in 2004, 2007, and 2011, respectively. He was a Research Fellow with the School of Computer Engineering and Institute of Media Innovation, Nanyang Technological University, Singapore. He is currently an Associate Professor with the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology. His research interests involve computer vision, image processing, and machine learning. Dr. Xiao was a recipient of the IEEE Innovation Spotlight Research Paper Award 2020, the EurAgEng Outstanding Paper Award 2018, and the Best Paper Award at ICIRA 2018. He also serves as the Associate Editor for *IET Image Processing*.

Joey Tianyi Zhou is a principal scientist, Investigator and group manager with A*STAR Centre for Frontier AI Research (CFAR), Singapore. He is also holding an adjunct faculty position at the National University of Singapore (NUS). Before working at CFAR, he was a senior research engineer with SONY US Research Center in San Jose, USA. Dr. Zhou received a Ph.D. degree in computer science from Nanyang Technological University (NTU), Singapore. His current interests mainly focus on improving the efficiency and robustness of machine learning algorithms. In these areas, he has published more than 100 papers and received the Best Student Paper Nomination at the European Conference on Computer Vision (ECCV'16), Best Paper Award at the International Joint Conference on Artificial Intelligence (IJCAI) workshops, and Best Poster Award and runner- up prize at International Conference on Computer Vision (ICCV19) on HANDS workshop and its competition, respectively. Dr. Zhou regularly organizes workshops/tutorials at top-tier international conferences like CVPR, IJCAI, ICDCS, etc. He is serving as an Associate Editor for IEEE Transactions on Emerging Topics in Computational Intelligence (TETCI) and IEEE Access, IET Image Processing, and Area Chairs in top machine learning conferences like ICLR, ICML, NeurIPS etc.

Zhiwen Fang received the B.S. and M.S. degrees from the Automation School, Beihang University, and the Ph.D. degree from the Huazhong University of Science and Technology, Wuhan, China, in 2004, 2008, and 2017, respectively. He was a Research Fellow with the Institute of Media Innovation, Nanyang Technological University, and a Research Scientist with the Institute of High Performance Computing, Research Agency for Science, Technology, and Research, Singapore. He is currently an Associate Professor with the School of Biomedical Engineering, Southern Medical University, Guangzhou, China. His research interests include medical image analysis, object detection, anomaly detection and machine learning. Dr. Fang also serves as an Associate Editor for *IET Image Processing*.

Feng Yang received his M.S. degree in biomedical signal and image processing from Sun Yat-Sen University, China in 1993, and the Ph.D. degree in communication and electronic systems from South China University of Technology, China in 1998. He joined the Division of Image Processing (LKEB) in Leiden University Medical Center, Netherlands, from April 2010 to April 2011 as a Visiting Scholar. He is currently with the School of Biomedical Engineering, Southern Medical University, Guangzhou, China, as a Professor and the Director in the Department of Electronic Technology. His research interests include wavelet analysis, medical image processing, and pattern recognition.