Pattern Recognition 143 (2023) 109776

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

# Classification of tumor in one single ultrasound image via a novel multi-view learning strategy

Yaozhong Luo[a,b], Qinghua Huang[a,b,*], Longzhong Liu[c]

[a] School of Information and Control Engineering, Xi'an University of Architecture and Technology, Xi'an 710055, China
[b] School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China
[c] Department of Ultrasound, The Cancer Center of Sun Yat-sen University, State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine, 510641, Guangdong, China

ARTICLE INFO

*Article history:*
Received 7 January 2023
Revised 11 April 2023
Accepted 21 June 2023
Available online 24 June 2023

*Keywords:*
Image classification
Deep learning
Multi-view learning
Breast cancer recognition

ABSTRACT

Computer-aided diagnosis (CAD) technology has been widely used in the early diagnosis of breast cancer. Nowadays, most of the existing breast ultrasound classification methods need to crop a tumor-centered image (TCI) on each image as the input of the system. These methods ignore the fact that the tumor as well as its surrounding tissues can actually be viewed from multiple aspects, and it is difficult to extract multi-resolution information applying only a single view image. In addition, the current methods do not effectively extract fine-grained features, and subtle details play an important role in breast classification. In our research, we propose a novel strategy to generate multi-resolution TCIs in a single ultrasound image, resulting in a multi-data-input learning task. Hence, a conventional single image based learning task is converted into a multi-view learning task, and an improved combined style fusion method suitable for a deep network is proposed, which integrates the advantage of the decision-based and feature-based methods to fuse the information of different views. At the same time, we first attempt to introduce the fine-grained classification method into breast classifications and capture the pairwise correlation between feature channels at each position to extract subtle information. The comparative experimental results show that our method can effectively improve the classification performance and achieves the best results in five metrics.

## 1. Introduction

Breast cancer is a severe disease and accounts for nearly a quarter of all cancer cases in women worldwide [1]. Considering the complex etiology of breast cancer, early detection and treatment are the key measures to reduce mortality [2]. In recent years, medical imaging technology has become the mainstream scheme of primary screening of breast tumors. Among them, ultrasound is one of the most popular methods because of the advantages of low cost, no radiation, faster imaging, and increasing accuracy and sensitivity [3].

However, because of its imaging characteristics, ultrasound images are more difficult to understand and require well-trained radiologists to make an accurate diagnosis [4]. Therefore, in order to help clinicians make a more accurate diagnosis, a large number of researchers have developed a series of computer-aided diagnosis (CAD) systems using machine learning methods [5]. The CAD system can provide quantitative support for clinical decision-making by effectively analyzing a large number of images, and their clinical application value has been confirmed in the actual diagnosis, which is helpful for radiologists to classify and identify breast tumors [6]. Traditional classification methods mainly include two aspects: feature extraction and classification [7,8]. The two steps are designed separately and then integrated to achieve the overall performance. Effective feature extraction is the key to breast tumor classification, mainly including texture features [9,10] and morphological features [11]. However, these features are often shallow and low-level, which is difficult to comprehensively describe the distinguishing patterns of tumors [12]. In addition, meaningful feature extraction is highly dependent on the quality of the preprocessing processes, such as segmentation, which usually requires recursive experiments to achieve satisfactory results. At the same time, the separation of feature extraction and classification is not conducive to improving the classification effect.

In recent years, deep convolution neural network (DCNN) methods have been concerned by researchers and used in breast cancer CAD systems because of their excellent feature learning ability [13,14]. In the research of breast ultrasound classification, the main

[*] Corresponding author.
*E-mail address:* qhhuang@nwpu.edu.cn (Q. Huang).

https://doi.org/10.1016/j.patcog.2023.109776
0031-3203/© 2023 Elsevier Ltd. All rights reserved.

deep learning based methods firstly automatically segment or ask the operators to provide a rectangular region of interest centered on the tumor [15], and the region is called the tumor-centered image (TCI). After the standardization of the TCIs, the DCNN is applied for feature extraction and classification. In [16], Xiao et al. asked the user to segment the TCIs, applied three pre-trained DCNNs, and fine-tuned them on the breast ultrasound dataset. Then the features extracted from three networks were combined, and an artificial neural network was adopted for the tumor classification. In [17], Liao et al. applied a block-based region segmentation algorithm to obtain the TCIs for the prediction of breast cancer. In [18], Cao et al. apply manually selected TCIs by clinicians to train DCNNs and analyze the performance of different DCNN methods based on the TCIs and full images. And the deep learning method based on TCIs obtained better performance than that based on full images. In [19], the TCIs containing breast lesions were provided and pre-processed based on image enhancement and bilateral filtering methods, respectively. Then the TCI and pre-processing images were combined as input of the DCNN for breast ultrasound lesion classification. In [20], Daoud et al. selected the TCI in the ultrasound image and resized it as the input layer of the AlexNet model for the breast tumors classification. In [21] and [22], Zeimarani et al. and Wang et al. first cropped the TCIs and applied the shallow networks to train and classify breast tumors.

However, the existing breast tumor classification algorithms based on deep learning still have some problems. On the one hand, the current breast ultrasound classification methods often need to crop a TCI on each ultrasound image as the input of the neural network, which is a single-view learning method. These methods ignore the fact that the tumor as well as its surrounding tissues can actually be viewed from multiple aspects, and it is difficult to extract multi-resolution information applying only a single view image as the input of the network. In addition, obtaining TCIs is empirical, and there is no standard. Different TCI partition methods have a significant impact on the final classification. At the same time, because the deep learning based classification method needs to resize the image to a uniform size, the size of the tumor region is also affected by the boundary. If the exterior margin is too small, or the TCI even only contains the tumor area, the edge infiltration features will be challenging to extract. If the boundary is too large, there will be more information, but the tumor region is with a smaller proportion, and the tumor details may be weakened in the network feature extraction. In addition, some artifacts and noises on the ultrasound image will also be included in the TCI. The optimal extended pixel value is different for each tumor. Therefore, different TCI partition methods have a great impact on the final classification [23]. At the same time, because different images have different requirements for TCI expansion ratio, it is difficult to find the best TCI acquisition method. On the other hand, there is a limited difference between benign and malignant breast tumors. Clinically, experts are even required to capture the decisive details and features, such as subtle calcification, the category of aftersound [24]. The characteristics that distinguish between benign and malignant tumors are often fine-grained, and the details of the ultrasound image greatly impact the identification of benign and malignant [25,26]. However, the existing breast ultrasound classification methods ignore the extraction of fine-grained features. In the existing deep learning method for breast ultrasound classification, after the convolution feature extraction layer, they often use a first-order processing method such as global average pooling to encode the features. However, this coding method is not conducive to the extraction of subtle features. Fine-grained learning has not been paid attention to the breast ultrasound classification. Therefore, it is necessary to explore how to add the fine-grained feature extraction method to the breast ultrasound classification network.

Therefore, we propose a novel multi-view homogeneous bilinear model for breast ultrasound classification. Different from the existing breast ultrasound classification method, we design a multi-resolution TCIs generation rule to obtain multiple TCIs for each image. Each TCI is regarded as a view of the tumor, and a multi-view learning method is applied to classify the breast image. In addition, to extract fine-grained features for breast tumor classification, we propose to apply a homogeneous bilinear network for the feature extraction of every view. The homogeneous bilinear network uses a bilinear pooling method to encode the features obtained from the convolution network based on correlation analysis, extract the second-order information of the feature layer, and receive more discriminative and fine-grained texture features. Therefore, the main contributions are summarized as follows:

(1) A novel classification scheme based on multi-view learning is proposed to extract multi-resolution information for single ultrasound image diagnosis. In this scheme, each image generates a TCI series with different resolutions, and each TCI is regarded as a view of the tumor. A conventional single image based learning task is converted into a multi-view learning task.
(2) Inspired by fine-grained recognition, to extract more representative features, we first attempt to apply fine-grained learning to breast ultrasound classification and propose to apply a homogeneous bilinear network for breast tumor feature extraction.
(3) We integrate the advantages of early and late fusion and propose a combined style fusion method suitable for deep networks. The fusion method aggregates the features and decisions of multiple views for final classification.
(4) The experimental results show that the multi-view homogeneous bilinear framework can improve breast cancer diagnosis performance.

This paper is organized as follows. Section 2 introduces the specific network structure and algorithm steps. Then, Section 3 presents the experimental results and the comparison between different methods. Finally, Section 4 discusses the research and gives the conclusion.

## 2. Methods

In this section, we introduce our breast ultrasound classification method based on multi-view learning and fine-grained recognition through a step-by-step presentation of our scheme and the detailed structure.

### 2.1. Overall architecture

Fig. 1 shows the specific steps of our proposed scheme. It mainly includes three procedures:

(1) Creating multiple TCIs with different resolutions

In order to extract the multi-resolution features of breast tumors and meet the needs of different tumors for different expansion rates, we first design a multi-resolution TCI group generation criterion, which generates TCI series from each breast ultrasound image for subsequent feature extraction. Each TCI is regarded as a view of the tumor. This module inputs a whole image and outputs a series of TCIs.

(2) The homogeneous bilinear network for each view

After obtaining the multi-resolution TCIs for each image, in order to extract fine-grained features, this study proposes a homologous bilinear network for each type of TCI view. The input of each network is the TCI corresponding to the view, and the output is the discrimination probability and feature representation.

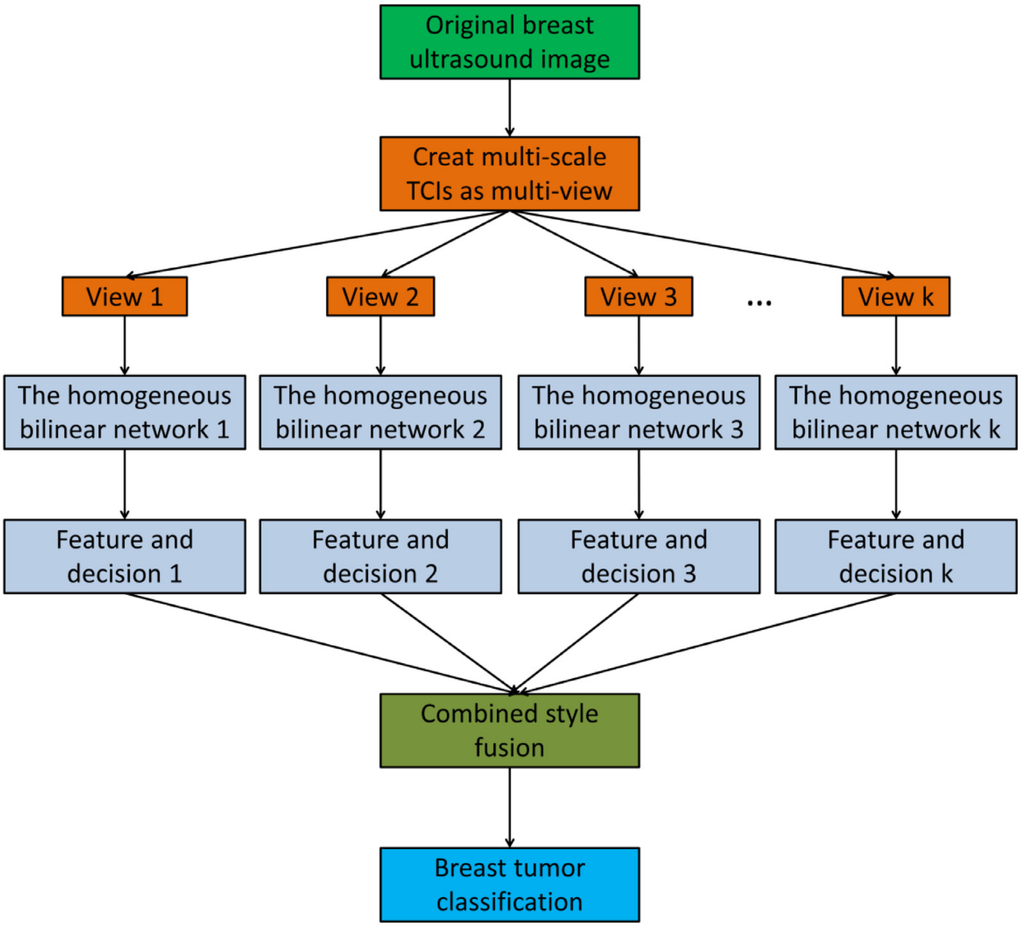
Original breast ultrasound image
Creat multi-scale TCIs as multi-view
View 1
View 2
View 3
...
View k
The homogeneous bilinear network 1
The homogeneous bilinear network 2
The homogeneous bilinear network 3
The homogeneous bilinear network k
Feature and decision 1
Feature and decision 2
Feature and decision 3
Feature and decision k
Combined style fusion
Breast tumor classification

**Fig. 1.** The flowchart of the proposed scheme. It consists of three steps: creating the multi-resolution TCI group, the homogeneous bilinear network for each view, and a combined style fusion module for different views.

(3) Combined style fusion module for different views

In order to fuse the decision and features of different views, we propose a combined style fusion method for the multiple networks. The combined style multi-view learning method combines the early and late fusion methods to get comprehensive results. The fusion module inputs the features extracted from the multi-view network and the discrimination probability of each view and outputs the category of breast tumor.

Our breast tumor classification scheme involves three problems: (1) How to avoid the difficulty of selecting the optimal TCI, not only input more information but also gather image different scale details and extract multi-resolution features conveniently. (2) How to extract finer-grained features and obtain more effective classification-related representation. (3) How to fuse the multiple networks of different views to get more practical features and classification results. Specific technologies will be introduced in the following sections.

### 2.2. Creating multiple TCIs with different resolutions

Previous deep learning based breast ultrasound classification methods usually need to crop a TCI as the input of the network. Because the infiltration of the tumor edge plays an essential role in the differentiation of benign and malignant tumors, researchers will segment the tumor region and expand it to get the TCI. As shown in Fig. 2, there is a distance between the actual leisure

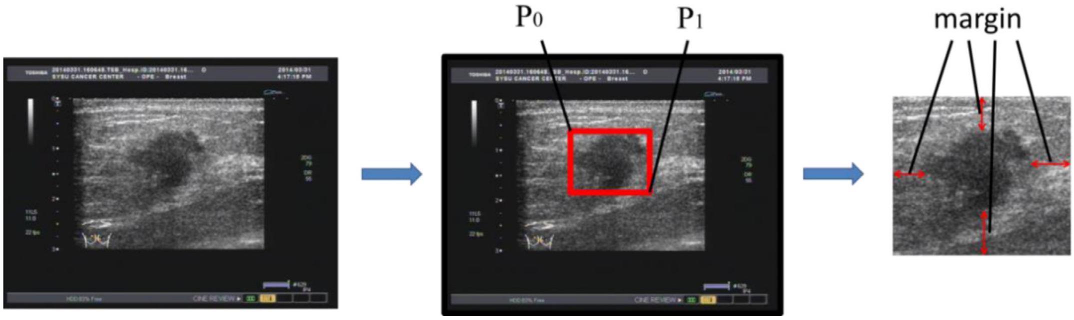

**Fig. 2.** The process of TCI extension. The image in the second column shows the tumor region and the seed points. The image in the third column shows the margin definition and the TCI obtained by one expansion.

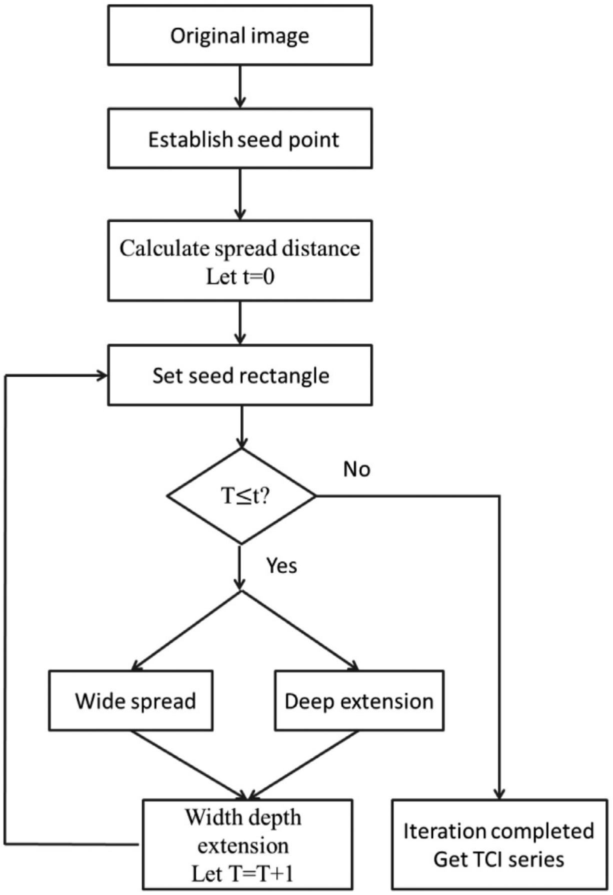

**Fig. 3.** The flowchart of the TCI series creation process.

boundary and the boundary of the crossed TCI itself. Therefore, the TCIs will include the lesion area and the background outside the lesion, which is convenient for extracting the information about the change from the edge to the outside, that is, the edge infiltration of the tumor. Because it is difficult to explore the optimal TCI and in order to obtain the information with different resolutions, we are not committed to finding an optimal TCI but propose a multi-resolution TCI acquisition approach for breast tumors classification. We design a certain rule and construct a set of TCIs with different outward expansions. The set includes multiple extended TCIs, and each TCI extends a certain proportion along the depth direction, width direction, or height and width direction concurrently outside the tumor boundary.

Suppose that R represents the whole image and the coordinates of the upper left and lower right points are $P_0 = (x_0, y_0)$ and $P_1 = (x_1, y_1)$. $R_{W0} : [x_0 : x_1, y_0 : y_1]$ represents the tumor region, and $w = x_1 - x_0$ and $h = y_1 - y_0$ represent the width and height of the tumor region, respectively. Fig. 3 shows the flowchart of the TCI series creation process. The seed points $P_0$ and $P_1$ are annotated by radiologists using Wada's LabelMe software [27]. We extend to the left and right sides with the width-margin w/m pixels and extend to the up and down with the height-margin h/m pixels. 1/m represents the expansion rate and in our study, m is set to 8. At the same time, to increase the contrasting diversity of the tumors with different heights and widths, a strategy of different height and width expansion is applied in the TCI acquisition method. As shown in Fig. 4, $R_{W2}$ extends in the width direction, $R_{W3}$ extends in the height direction, and $R_{W4}$ extends in both width and height direction. $R_{W2}$, $R_{W3}$, and $R_{W4}$are three TCIs obtained by one extension. Our method is to get a series of TCIs with a gradually increasing margin outside the tumor. The second extension is based on the first extension. The times of expansion are defined as t. The more times of expansion, the more diverse views of the tumor, but it will also bring more calculating expenses. In our study, the value of t was set to 2. The arrays can express all TCIs in Formula (1).

After that, all TCIs are standardized into unified 224 × 224. Hence, different views contain different outward expansions and correspond to diverse tumor scale sampling. When they input the neural network, multi-resolution features can be extracted. Each TCI is represented in the form of an array, as shown in the following formula, and the specific image is shown in Fig. 4.

$$\begin{aligned}
&R_{W0} : [x_0 : x_1, y_0 : y_1] \\
&R_{W1} : \left[x_0 : x_1, y_0 - \frac{y_1 - y_0}{8} : y_1 + \frac{y_1 - y_0}{8}\right] \\
&R_{W2} : \left[x_0 - \frac{x_1 - x_0}{8} : x_1 + \frac{x_1 - x_0}{8}, y_0 : y_1\right] \\
&R_{W3} : \left[x_0 - \frac{x_1 - x_0}{8} : x_1 + \frac{x_1 - x_0}{8}, y_0 - \frac{y_1 - y_0}{8} : y_1 + \frac{y_1 - y_0}{8}\right] \\
&R_{W4} : \left[x_0 - \frac{x_1 - x_0}{8} : x_1 + \frac{x_1 - x_0}{8}, y_0 - 2 \times \frac{y_1 - y_0}{8} : y_1 + 2 \times \frac{y_1 - y_0}{8}\right] \\
&R_{W5} : \left[x_0 - 2 \times \frac{x_1 - x_0}{8} : x_1 + 2 \times \frac{x_1 - x_0}{8}, y_0 - \frac{y_1 - y_0}{8} : y_1 + \frac{y_1 - y_0}{8}\right] \\
&R_{W6} : \left[x_0 - 2 \times \frac{x_1 - x_0}{8} : x_1 + 2 \times \frac{x_1 - x_0}{8}, y_0 - 2 \times \frac{y_1 - y_0}{8} : y_1 + 2 \times \frac{y_1 - y_0}{8}\right]
\end{aligned} \tag{1}$$

Taking t = 2 as an example, seven different TCIs have different region expansions, corresponding to the needs of different edge infiltration information. At the same time, since all TCIs are standardized into 224 × 224 images as the input of the different networks, the relative proportion of tumors is different. Corresponding to different sampling rates, the receptive field of each neuron is different. Therefore, our method can extract multi-resolution features by inputting neural networks with varying sampling rates.

### *2.3. The homogeneous bilinear network for each view*

After obtaining the multi-resolution TCI group for each image, we regard each TCI as a view of the breast tumor and train a homologous bilinear network for each type of TCI view. In our study, seven networks correspond to the number of TCI types. In [28], Lin et al. theoretically proved that various orderless texture descriptors can be written in the bilinear form and can effectively extract fine-grained features through end-to-end training. Hence, in order to extract more fine-grained features of the breast tumor, we propose to apply a homogeneous bilinear network for breast ultrasound classification. In the proposed homogeneous bilinear network, the texture feature extraction of breast tumor image is regarded as a bilinear form of a homogeneous convolutional neural network. Unlike the classical ultrasound classification network, which implements first-order information from the feature output of the backbone, our network focuses on extracting second-order information from features extracted by the backbone network. The homogeneous bilinear network consists of three components: the homogeneous backbone network, the bilinear module, and the classification layer.

Because the Xception network has a powerful feature representation ability, it is selected as the backbone network in this study. Xception [29] replaces standard convolution with a depthwise separable convolution layer. Depthwise separable convolution performs convolution operations on spatial and channel dimensions successively, which reduces the number of weight coefficients while preserving the representation learning ability of convolution kernels. As shown in Fig. 5, the input of our Xcetpion is the TCI, and the output of the last separable convolution layer with ReLU activation is the feature representation. More technically, let $\boldsymbol{\mu_{ori}}$ be the deep feature representation with shape (7, 7, 2048),

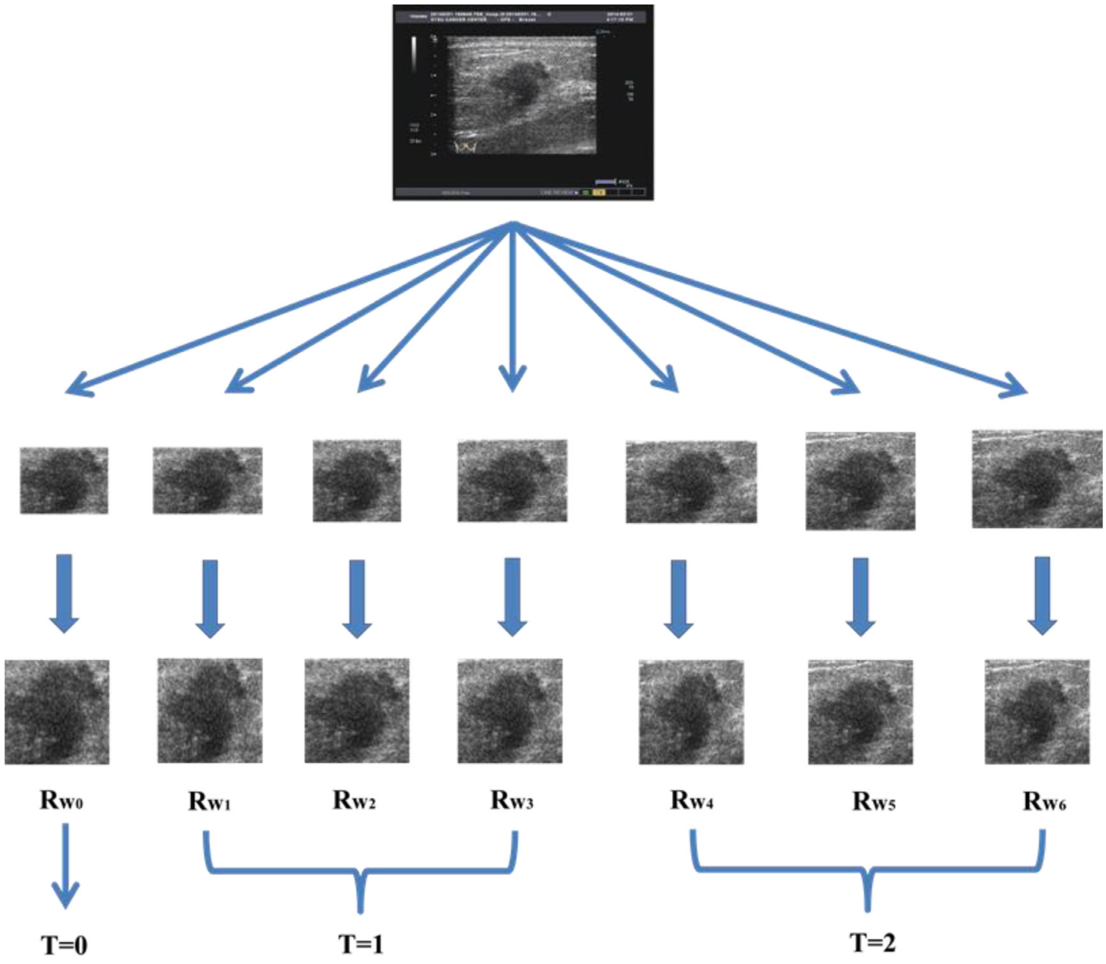

**Fig. 4.** Generation and standardization of TCIs with different resolutions. The image in the first line is the original image, the second line is the cropped TCI with varying rates of expansion, and the third line is the result after the standardization. Through this process, we have obtained TCIs with different resolutions.

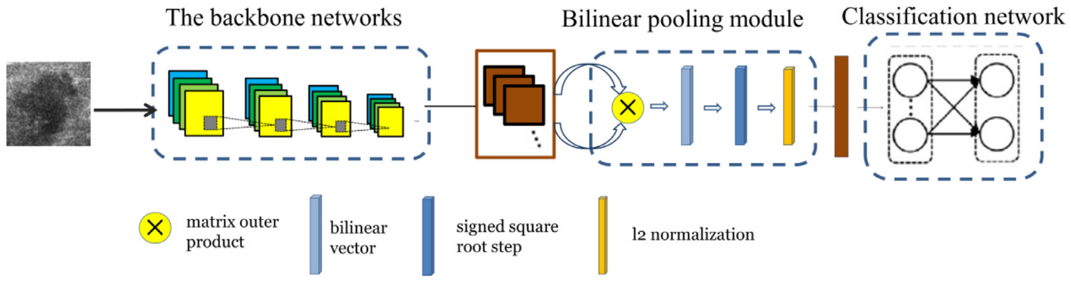

**Fig. 5.** The specific network structure of the homologous bilinear network. The homogeneous bilinear network consists of three components: the homogeneous backbone network, the bilinear pooling module, and the classification network.

and each value in (7, 7) represents the high-order representation of a local position. The homogeneous feature extraction network applies one CNN to extract features and obtains the high-level feature of different positions.

After that, the bilinear pooling module is applied to calculate the second-order statistics of the feature activations extracted by the backbone networks to capture the relationship between the features of each position and generate the expressive global representation. The bilinear feature module includes two steps. Firstly, reshape the feature map into a two-dimensional matrix. Given the $\mathrm{K_{th}}$ TCI $\boldsymbol{I_K}$, $\boldsymbol{F_k}$ refer to the feature map extracted by the Xception backbone for the $\mathrm{K_{th}}$ TCI, the feature map and the reshape process can be formulated as follows:

$$\mathbf{F_K} = f_{Xception,K}(\mathbf{W_K}; \mathbf{I_K}) \tag{2}$$

$$\mathbf{V_K} = \mathrm{Reshape}(\mathbf{F_K}) \tag{3}$$

where $f_{Xception,K}$ refers to the Xception backbone of the $\mathrm{K_{th}}$ TCI and $\mathbf{V_K}$ is the result of the reshaping of $\mathbf{F_K}$ in terms of the third mode. Hence, $\mathbf{F_K} \in \mathbb{R}^{7\times7\times2048}$ and $\mathbf{V_K} \in \mathbb{R}^{49\times2048}$. Secondly, calculate the autocorrelation matrix of the multi-channel feature on each location, realize it through a bilinear pooling and get the feature combination matrix. The bilinear feature is represented as $\mathrm{bilinear}(L, \mathbf{I}, \mathbf{V_K}, \mathbf{V_K}) = \mathbf{V}_K(L, \mathbf{I})^{\mathrm{T}}\mathbf{V}_K(L, \mathbf{I})$. Where location L covers location and scale, and I refer to the image. This process is equivalent to calculating the product of feature activations of each position and then introducing a sum-pooling method. We computed the outer product for each local feature, then passed through the signed square root step (4), followed by $\mathrm{L_2}$ normalization (5), and got the general image descriptor. The bilinear feature for the $\mathrm{K_{th}}$

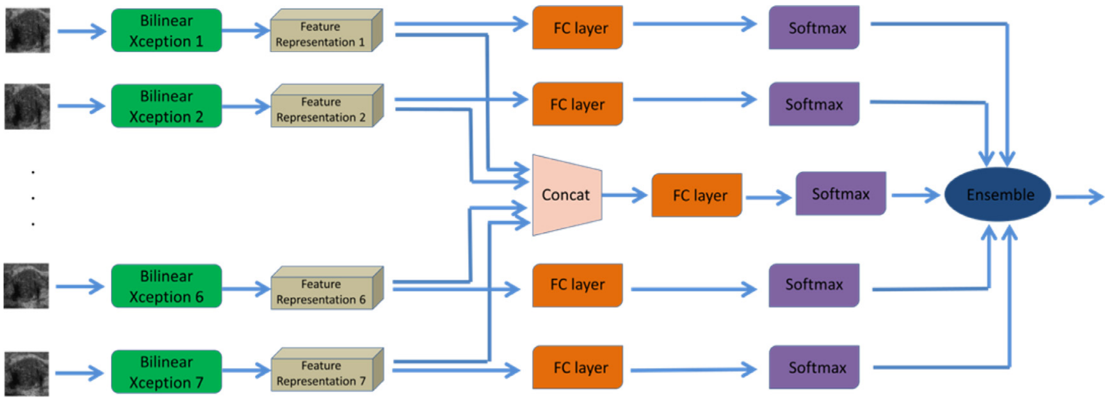

**Fig. 6.** The architecture of the combined style fusion method for multi-view networks. By aggregating the features of multiple views and training a recognizer, the discrimination probability of this recognizer is weighted with the discrimination probability of the networks from every single view.

TCI can be formulated as follows:

$$\mathbf{S_K} = \mathbf{V_K}\mathbf{V_K^T} \tag{4}$$

$$\mathbf{X_K} = \mathrm{sign}(\mathbf{S_K})\sqrt{(|\mathbf{S_K}|)} \tag{5}$$

$$\mathbf{Y_K} = \mathbf{X_K}/ \parallel \mathbf{X_K} \parallel^2 \tag{6}$$

$$\mathbf{Z_K} = \mathrm{Dropout}(\mathbf{Y_K}) \tag{7}$$

$$D_K = \mathrm{f}_{softmax}(\mathbf{W_{FC,K}}\mathbf{Z_K}) = \delta(\mathbf{W_{FC,K}}\mathbf{Z_K}) \tag{8}$$

where $\mathbf{S_K}$ is the result of the outer product of the feature matrix and then sum-pooling over 49 spatial locations to produce a holistic representation of the image with dimension $2048^2$, signis the signum function, and the rate of Dropout is set to 0.5. In our study, the discrimination result and feature representation are obtained from each network. As shown in Fig. 4, $D_K$ is the decision result, while the feature output before the fully connected layer $\mathbf{Z_K}$ is the feature representation.

The homogeneous bilinear network examines the interaction between different channels by calculating the outer product of the convolution descriptor. The outer product captures the pairwise correlation between feature channels at each position, which is translation invariant. Because different semantic features are extracted from different channels of the description vector, the relationship between different semantic features of the input image can be captured at the same time through the second-order pooling operation. It provides more fine-grained feature representation than first-order pooling. Because the location dimension of the feature is pooled, the obtained bilinear feature is orderless.

### *2.4. Combined style fusion method for multi-view network*

To perform a comprehensive diagnosis based on different views, we use a multi-view fusion method to combine the information from various networks. There are two kinds of fusion methods. One is early fusion through the combination of features, and the other is late fusion through the combination of decision results [30,31]. In our research, we integrate the advantages of the two kinds of fusion methods and propose a combined-style method suitable for the deep network. In our combined style fusion method, by aggregating the features of multiple views and training a recognizer, the discrimination probability of this recognizer is weighted with the discrimination probability of the networks from every single view. The multi-view fusion process is shown in Fig. 6.

Firstly, the feature output of the layer in front of the softmax layer of the seven models corresponding to seven views is concatenated as the aggregated feature. Given the extracted feature representation of seven networks $\mathbf{Z_{K0}}$, $\mathbf{Z_{K1}}$, ..., $\mathbf{Z_{K6}}$, the aggregated feature for a breast ultrasound image can be formulated as follows:

$$\mathbf{Z_{agg}} = \mathrm{concatenate}(\mathbf{Z_{K0}}, \mathbf{Z_{K1}}, ..., \mathbf{Z_{K6}}) \tag{9}$$

where *concatenate* is an operator which concatenates multiple tensors into a tensor. Then, a classification network for the aggregated feature is applied. It mainly includes a fully connected layer with a softmax activation, which is the same as the network applying for each view. Because different networks correspond to different views and each view contains different extension contents and is with different resolutions, this process combines the complementary and redundant features extracted by each network through the fully connected layer, which can obtain a more robust recognizer. The output of the softmax layer is regarded as the discrimination decision probability of the aggregated recognizer and can be formulated as follows:

$$\boldsymbol{D_{agg}} = f_{softmax}\left(\boldsymbol{W_{FC,agg}}\boldsymbol{Z_{agg}}\right) = \delta\left(\boldsymbol{W_{FC,agg}}\boldsymbol{Z_{agg}}\right) \tag{10}$$

Secondly, the discrimination probability of aggregated recognizers is weighted with the discrimination probability of the networks from every single view. This process can be regarded as the combination of a feature fusion based recognizer and the traditional result based fusion method. The mathematical process of this process can be expressed by the following formula:

$$\boldsymbol{D_{final}} = \ \mathrm{a} \ \times \boldsymbol{D_{agg}} + \mathrm{b} \ \times \frac{1}{N_T}\sum_T \boldsymbol{D_T} \tag{11}$$

Because feature-based fusion and result-based fusion have almost similar fusion performance, we give them the same weight, setting a = b = 1 / 2. $\boldsymbol{D_{final}}$ refers to the final decision which is given by combined style fusion of multiple networks.

### *2.5. Training process*

Firstly, multi-resolution TCIs with different expansion rates are generated for each image on the breast ultrasound dataset. Then the homologous bilinear networks are trained for each type of TCI. In the training process of each single view network, our study applies the Adam algorithm to update the parameters and sets the initial learning rate to 0.0001. In addition, the binary cross-entropy is applied to evaluate the distance between the predicted value and the true value. When training each single view homologous bilinear network, data augmentation methods such as rotation, flip, zoom, shearing, and shift are used. After training the single view homologous bilinear network, this study aggregates the features of each view and trains the aggregated feature recognizer. Finally, the discrimination probability of the aggregated feature recognizer is weighted with the discrimination probability of the networks from every single view.

**Table 1**
The cardinality of Dataset.

| Dataset | Patient | Image |
|---|---|---|
| **Benign** | 291 | 786 |
| **Malignant** | 305 | 916 |
| **Total** | 596 | 1702 |

### *2.6. Experimental methods*

#### *2.6.1. Data description*

1702 ultrasound images from 596 patients were provided by the Cancer Center of Sun Yat-sen University and used in our experiment. All images are labeled with physician pathology results. Due to the focus of this research on breast tumor classification, the dataset only includes cases of breast tumors, while images of normal cases without tumors are screened by radiologists and not included in the experimental dataset. In addition, as shown in Fig. 4, although colored imprints and logos were added to the images during the data collection process, all cropped TCIs will be converted into grayscale images for subsequent analysis because ultrasound images are originally grayscale images. Table 1 shows the data set used for our experiment, including the number of benign and malignant patients and images.

#### *2.6.2. Comparative experiment*

To verify our proposed multi-view fine-grained based breast cancer diagnosis method, we design a five-fold cross-validation experiment. We divide the breast ultrasound image dataset into five parts. In addition, we put the images of the same patient into the same fold to avoid the images of the same case appearing in the training set and the test set at the same time, which will affect the experimental accuracy [38]. After averaging the results of five experiments, they are used to describe the classification performance of the algorithms [32].

Moreover, we design three kinds of experiments on the breast dataset. The first is the comparative experiment with the existing breast ultrasound classification methods. The proposed method is compared with the existing eleven breast image classification methods based on deep learning. These eleven methods are Daoud et al. [20], Xiao et al. [16], Cao et al. [18], Zhuang et al. [19], Zeimarani et al. [21], Wang et al. [22], Hijab et al. [33], Tanaka et al. [34], Zhuang et al. [36], Kim et al. [35], Luo et al. [37], and the baseline Xception. These works first empirically select a single view of the breast image and then apply the deep networks for the classification of breast tumors. In addition, in terms of the used deep network, the works in [20,33], and [18] are based on pre-trained AlexNet, VGG16, and DenseNet, respectively. The method in [16] is the feature combination model of three pre-trained networks and the work in [19] combined the original image and their preprocessing images as the input of pre-trained VGG16. The work in [34] combines the decision information of VGG19 and ResNet152 for breast tumor diagnosis. The methods in [22] and [21] are customized shallow networks whose numbers of the convolutional layer are only a few. In [35], the authors generated entropy and phase images from ultrasound images and classified breast tumors by combining three different feature maps. The work in [36] is based on the combination of image decomposition and adaptive spatial feature fusion for tumor classification. The work in [37] proposed a segmentation-to-classification framework that enhances the clinical-related features through two parallel feature branches for breast ultrasound classification. These algorithms are the works of breast tumor classification in recent years and have achieved good classification results. The second step is the ablation experiment of the proposed method. The ablation experiment is to demonstrate the effectiveness of multi-view generation, the homogeneous bilinear network, and the combined style fusion method. The third step is to compare the effects of different amplification parameters t. In order to facilitate comparison, all experiments were developed under the same hardware environment and software system.

#### *2.6.3. Evaluation metrics*

In order to evaluate the effect of different algorithms in breast tumor classification, we have used five metrics, including accuracy, sensitivity, specificity, $F_1$-score, and AUC, which are widely used in the evaluation of medical image classification [39].

## 3. Results

### *3.1. Comparison with existing methods*

We compared our multi-view learning method with the breast classification methods in recent years, and the experimental results are shown in Table 2. These methods select one view from each ultrasound image as the input of the DCNN and train the networks for breast tumor classification. Among them, although the custom networks [21,22] have few training parameters, the performance is not as good as the transfer learning methods [16,18,20,33,34]. Moreover, our method has also achieved the best results, with the highest accuracy (92.12%), sensitivity (95.31%), specificity (88.41%), $F_1$-score (93.03%), and AUC (0.9743). Higher AUC and accuracy are obtained, which shows better comprehensive classification performance. Our method has received the highest sensitivity and can better detect malignant tumors. For breast tumor classification, the improvement of sensitivity is more gratifying and valued. In addition, the improvement of the $F_1$-score (93.03%) means that the proposed scheme can better balance the relationship between sensitivity and specificity in classification. The highest effect was obtained in all five metrics, which illustrates the proposed multi-view homogeneous bilinear network, which adds multi-view learning and fine-grained feature extraction to the classification network and can better distinguish benign and malignant tumors.

Furthermore, in order to more intuitively demonstrate the performance of our proposed method in breast tumor classification, we have shown the ROC curves and bar charts of different classification methods in Figs. 7 and 8, respectively. The ROC curve of the proposed method is higher than that of all comparative algorithms, and our research has achieved the highest classification accuracy in both benign and malignant tumor samples. They further prove the effectiveness of our proposed method.

### *3.2. Ablation experiment*

To prove the effectiveness of the proposed multi-view learning framework and the homogeneous bilinear network, we conducted ablation experiments. Since this study comprises three essential steps, the ablation experiment will be discussed from these aspects.

#### *3.2.1. Effectiveness of multi-resolution TCIs generation strategy*

Firstly, to prove the effectiveness of the multi-resolution TCIs generation strategy for breast tumor classification, we compare the proposed method with the method Proposed_WI, Proposed_Tumor, and Proposed_TCI, and the results are shown in Table 3. These

**Table 2**
Comparison result with other methods.

| Method | Accuracy | Sensitivity | Specificity | $F_1$-score | AUC |
|---|---|---|---|---|---|
| Daoud et al. [20] | 0.8219 | 0.8637 | 0.7734 | 0.8400 | 0.8976 |
| Xiao et al. [16] | 0.8407 | 0.8876 | 0.7861 | 0.8588 | 0.9238 |
| Cao et al. [18] | 0.8237 | 0.8581 | 0.7836 | 0.8413 | 0.9100 |
| Zhuang et al. [19] | 0.8214 | 0.8626 | 0.7735 | 0.8381 | 0.8988 |
| Zeimarani et al. [21] | 0.7914 | 0.8440 | 0.7301 | 0.8138 | 0.8784 |
| Wang et al. [22] | 0.7626 | 0.8734 | 0.6335 | 0.7996 | 0.8207 |
| Hijab et al. [33] | 0.8114 | 0.8593 | 0.7556 | 0.8305 | 0.8911 |
| Tanaka et al. [34] | 0.8255 | 0.8746 | 0.7683 | 0.8445 | 0.9152 |
| Kim et al. [35] | 0.8296 | 0.8996 | 0.7479 | 0.8540 | 0.9190 |
| Zhuang et al. [36] | 0.8431 | 0.8702 | 0.8116 | 0.8564 | 0.9142 |
| Luo et al. [37] | 0.8548 | 0.8691 | 0.8383 | 0.8661 | 0.9207 |
| Xception | 0.8319 | 0.8942 | 0.7594 | 0.8527 | 0.9104 |
| Proposed method | **0.9212** | **0.9531** | **0.8841** | **0.9303** | **0.9743** |

**Table 3**
Ablation experimental results of different style of network inputs.

| Method | Accuracy | Sensitivity | Specificity | $F_1$-score | AUC |
|---|---|---|---|---|---|
| Proposed_WI | 0.8143 | 0.8723 | 0.7467 | 0.8368 | 0.9004 |
| Proposed_Tumor | 0.8290 | 0.8669 | 0.7849 | 0.8456 | 0.9081 |
| Proposed_TCI | 0.8396 | 0.8833 | 0.7886 | 0.8571 | 0.9244 |
| Proposed method | **0.9212** | **0.9531** | **0.8841** | **0.9303** | **0.9743** |

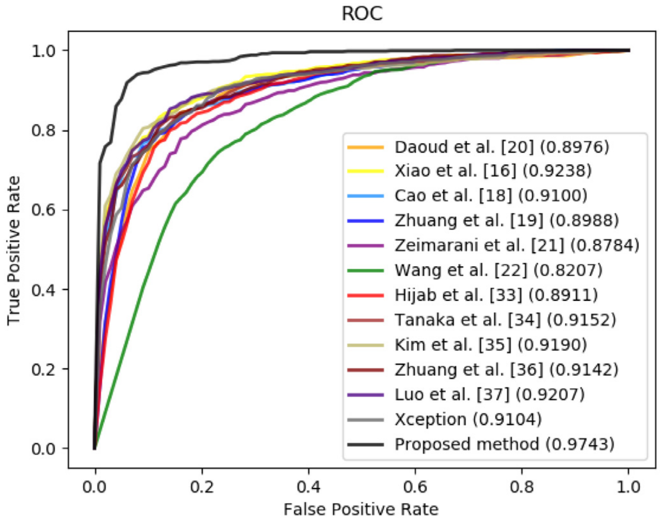

**Fig. 7.** ROC curves of different classification methods on breast ultrasound dataset.

methods are consistent with the proposed method in that the homologous bilinear network is still retained for breast tumor classification, but single-view learning methods are used. Among them, the method Proposed_WI applies the whole breast ultrasound image as the input of the network, Proposed_Tumor crops a rectangular region that covers the breast tumor as the input, and the method Proposed_TCI is to intercept a single TCI on the ultrasound image as the input of the network.

As shown in Table 3, the performance of the classification method using the whole ultrasound image and the rectangle just covering the tumor area is not as good as that using TCI. When only the rectangle region just covering the tumor is used as the input of the network, the information outside the tumor, including edge changes, infiltration, and other information, is not adequate. When using the whole image, although the information is more abundant, the ultrasound image is full of various noises, which will interfere with the classification of tumors. At the same time, the proportion of the tumor region in the whole image also becomes smaller. Therefore, the method which expands a certain distance outside the tumor and intercepts a TCI as network input can achieve better results than inputting the full image and the rectangle region just covering the tumor. In addition, our multi-

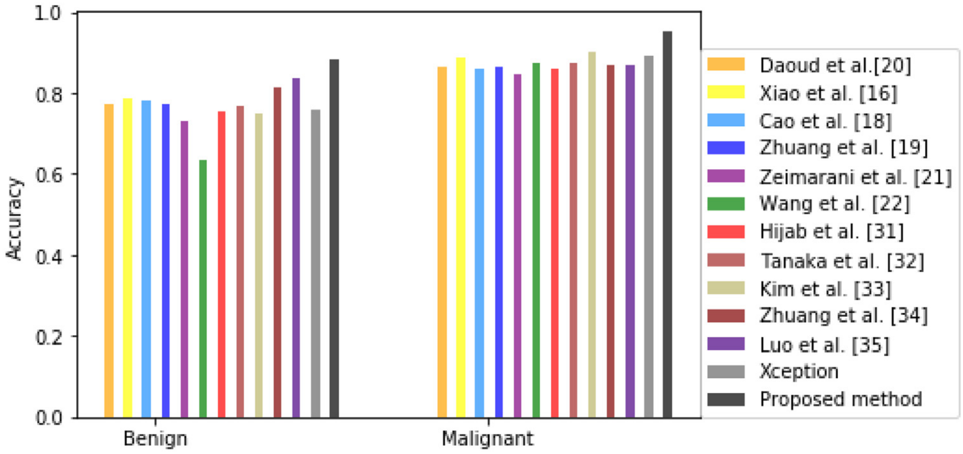

**Fig. 8.** Classification analysis for breast tumors classification based on different methods.

**Table 4**
Ablation experimental results of different multi-view fusion methods.

| Method | Accuracy | Sensitivity | Specificity | $F_1$-score | AUC |
|---|---|---|---|---|---|
| Proposed_TCI | 0.8396 | 0.8833 | 0.7886 | 0.8571 | 0.9244 |
| Proposed_DA | 0.8246 | 0.8750 | 0.7659 | 0.8447 | 0.9141 |
| Proposed_Concat | 0.8572 | 0.8734 | 0.8383 | 0.8689 | 0.9331 |
| Proposed_ADD | 0.8625 | 0.8778 | 0.8447 | 0.8736 | 0.9329 |
| Proposed_Voting | 0.8537 | 0.8953 | 0.8052 | 0.8693 | —— |
| Proposed_Average | 0.8513 | 0.8953 | 0.8001 | 0.8676 | 0.9300 |
| Proposed method | **0.9212** | **0.9531** | **0.8841** | **0.9303** | **0.9743** |

**Table 5**
Ablation experimental results of the homogeneous bilinear network.

| Method | Accuracy | Sensitivity | Specificity | $F_1$-score | AUC |
|---|---|---|---|---|---|
| Proposed_GP | 0.9018 | 0.9379 | 0.8599 | 0.9134 | 0.9702 |
| Proposed method | **0.9212** | **0.9531** | **0.8841** | **0.9303** | **0.9743** |

**Table 6**
Experiment results of different expansion times.

| Extended parameters | Accuracy | Sensitivity | Specificity | $F_1$-score | AUC |
|---|---|---|---|---|---|
| t=0 | 0.8290 | 0.8669 | 0.7849 | 0.8456 | 0.9081 |
| t=1 | 0.9124 | 0.9488 | 0.8701 | 0.9226 | 0.9702 |
| t=2 | **0.9212** | 0.9531 | **0.8841** | **0.9303** | 0.9743 |
| t=3 | 0.9148 | **0.9542** | 0.8688 | 0.9249 | 0.9734 |
| t=4 | 0.9159 | 0.9488 | 0.8777 | 0.9256 | **0.9751** |

view learning method achieves much better results than single-view learning and improves the accuracy and AUC by at least 8%. Because there is useful information for classification outside the tumor edge and the change of edge is useful for classification, constructing a certain rule to obtain a set of TCIs and taking them as the views of the breast tumor for multi-view learning can obtain different edge information and extract features at different scales, so as to obtain more robust and practical features for breast tumors classification. Our method is more effective than exploring a view of the tumor for classification. Because the series of TCI acquisition includes not only richer information but also the change of receptive field of each DCNN neuron corresponds to the extraction of multi-resolution information due to different sampling rates.

### *3.2.2. Effectiveness of the combined style multi-view fusion*

To verify the effectiveness of the combined style multi-view fusion strategy, we compared our method with the methods Proposed_DA, Proposed_Concat, Proposed_ADD, Proposed_Voting, and Proposed_Average, and the results are shown in Table 4. Proposed_DA does not use multi-view fusion methods but regards multi-resolution TCI generation as a data augmentation method for final classification. Methods Proposed_Concat and Proposed_ADD are feature-based multi-view learning methods. The former concatenates the features learned from seven views, and the latter adds the features learned from seven views for the breast tumor classification. In addition, Methods Proposed_Voting and Proposed_Average are decision-based multi-view learning methods, which average and vote the seven network decision information for breast tumor classification, respectively. It can be seen that the five multi-view learning methods are superior to the single-view learning methods, whether they are directly using TCIs or using multi-resolution generation as data augmentation. That means that applying the multi-resolution TCIs generation and multi-view learning is effective. In addition, our method achieves better results than these six methods, proving that the proposed fusion method combining the advantages of early fusion and late fusion is a more effective method for breast tumor classification.

### *3.2.3. Effectiveness of the homogeneous bilinear network*

Finally, to prove the effectiveness of the homogeneous bilinear network, we compare it with the method Proposed_GP, and the results are shown in Table 5. Methods Proposed_GP still used multi-view learning and replaced the bilinear pooling modules with the first-order processing operator global average pooling. Our heterologous bilinear network is better than Proposed_GP in five metrics and achieves accuracy and $F_1$-score improvement of 1.94% and 1.69%, respectively. It proves that extracting the second-order statistics after the backbone network can extract fine-grained features better than the first-order processing operator, which is helpful for breast ultrasound classification.

## *3.3. Experiment with different expansion times*

In addition, we evaluated the effect of different expansion times t on the proposed breast ultrasound classification scheme. Larger t means that more expanded TCIs are obtained into multi-view learning methods. The experimental results are shown in Table 6. From the comparison results, when t changes from 0 to 1, the effect is significantly improved, with a nearly 8% improvement. When t varies from 1 to 2, the performance also improves, but the progress begins to slow down. When t changes from 2 to 3 or 4, the performance improvements have almost stopped.

These results show that the increase in the number of extensions is not infinitely effective in classifying breast tumors. It is consistent with our intuitive understanding. When the selection of TCI is too large, the increased information is not beneficial to the improvement of classification, and even the irrelevant information is unfavorable to classification. As the expansion times increase, the calculation cost will increase. Since the calculation cost will increase with the increase of expansion times, t = 2 is the best choice for both performance and efficiency.

## **4. Discussion and conclusions**

Many researchers focus on breast tumor classification based on hand-crafted features and deep learning methods [40]. However, they need more attention to the problem of cropping TCI, and there is a lack of attention to fine-grained feature extraction in breast tumor feature extraction. The uniqueness of our work is to design a specific rule to construct multi-resolution TCIs with different outward expansions, apply a homogeneous bilinear network to calculate the correlation between feature maps and extract the fine-grained features, and propose a combined style method to fuse different views.

From the quantitative experimental results and ROC curves, compared with the breast tumor classification algorithms in recent years, our proposed methods have achieved the best results. The accuracy was improved by at least 6.64% and the $F_1$-score was improved by at least 6.42%. In addition, our algorithm has the highest sensitivity, which significantly reduces the missed diagnosis rate of malignant tumors. The experimental results show that our multi-view homogeneous bilinear model can extract tumor patterns more effectively and significantly improve classification performance.

In addition, as shown in Table 3, the effect of the multi-view learning method is better than the methods with the single view. These methods, whether based on the whole ultrasound image, the rectangle region just covering the tumor, or TCI, are not as good as the proposed method. Because they only take one view of the tumor, the information input to the network is limited. The multi-view method can obtain TCIs with different resolutions, which broadens the information representation of tumors. And the classification performance is effectively improved through the multi-view fusion method.

At the same time, the comparison with Proposed_GP also shows the effectiveness of our fine-grained feature extraction network. The classical method based on the deep network can extract image features, but it does not pay attention to the extraction of fine-grained features. However, the complexity of the medical morphology of ultrasound tumors leads to many classification differences in very small areas, such as calcification. The bilinear network simulates the multi-path method of human visual processing, and through bilinear operation, the relationship between different semantic features of the input image can be captured at the same time to extract more fine-grained features. The comparison proves the effectiveness of our homogeneous bilinear method. Moreover, the comparison with early and late fusion methods also shows the effectiveness of the combined-style fusion in the multi-view network. The feature-based aggregation method can integrate the features of different views for classification, and the further combination of the decision-making result of the feature aggregation network and other sub-networks can effectively improve the classification performance of the network.

Besides, we analyze the impact of different expansion times on the classification effect. The effect is significantly improved when we expand the first two times and begins to slow down on the subsequent expansion. It shows that the expansion of TCIs is not infinitely effective in the classification of breast tumors. When the selection of TCI is too large, the increased information is very small for the improvement of classification, and even the introduction of more irrelevant information is even unfavorable to classification.

Moreover, we discuss the limitations of our work. Firstly, our method requires the user to roughly give the upper left and lower right points of the tumor area. This step involves user participation rather than full automation. Meanwhile, inaccurate annotations may have a negative impact on classification results. And each doctor also has their own subjectivity when labeling. Consistently, how to automatically detect tumor areas is our subsequent work. Secondly, the training is time-consuming. Compared with a single network, our method includes networks from several views, which requires more calculation time and computing resources. Therefore, our next research will be to improve the reuse of the network and reduce the time cost of training. Finally, the analysis of breast ultrasound tumors not only distinguishes between benign and malignant, but also requires the classification of multiple risk levels. Therefore, applying the proposed method to solve multi-class classification is also one of the subsequent works.

Breast cancer is a serious disease that is harmful to women. The proposed multi-view learning method can effectively provide a second opinion for clinical diagnosis. At the same time, since the extraction of subtle features and effective view acquisition are common problems, the multi-view fine-grained analysis method is also of great significance for other disease diagnoses.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

Data will be made available on request.

## Acknowledgments

This work was supported in part by the National Key Research and Development Program under Grant 2018AAA0102100, the National Natural Science Foundation of China under Grants 62071382 and 82030047, and the Innovation Capability Support Program of Shaanxi under Program 021TD-57.

## References

[1] G. Li, C. An, J. Yu, Q. Huang, Radiomics analysis of ultrasonic image predicts sensitive effects of microwave ablation in treatment of patient with benign breast tumors, Biomed. Signal. Proces. 76 (2022) 103722, doi:10.1016/j.bspc.2022.103722.

[2] J. Song, Y. Zheng, J. Wang, M.Z. Ullah, X. Li, Z. Zou, G. Ding, Multi-feature deep information bottleneck network for breast cancer classification in contrast enhanced spectral mammography, Pattern Recogn. 131 (2022) 108858, doi:10.1016/j.patcog.2022.108858.

[3] J. Xi, D. Sun, C. Chang, S. Zhou, Q. Huang, An omics-to-omics joint knowledge association subtensor model for radiogenomics cross-modal modules from genomics and ultrasonic images of breast cancers, Comput. Biol. Med. 155 (2023) 106672, doi:10.1016/j.compbiomed.2023.106672.

[4] N. Karunanayake, W. Lohitvisate, S.S. Makhanov, Artificial life for segmentation of fusion ultrasound images of breast abnormalities, Pattern Recogn. 131 (2022) 108838, doi:10.1016/j.patcog.2022.108838.

[5] X. Fei, S. Zhou, X. Han, J. Wang, S. Ying, C. Chang, W. Zhou, J. Shi, Doubly supervised parameter transfer classifier for diagnosis of breast cancer with imbalanced ultrasound imaging modalities, Pattern Recogn. 120 (2021) 108139, doi:10.1016/j.patcog.2021.108139.

[6] D. Song, Z. Zhang, W. Li, L. Yuan, W. Zhang, Judgment of benign and early malignant colorectal tumors from ultrasound images with deep multi-view fusion, Comput. Methods Programs Biomed. 215 (2022) 106634, doi:10.1016/j.cmpb.2022.106634.

[7] S H Wang, D R Nayak, D S Guttery, et al., COVID-19 classification by CCSHNet with deep fusion using transfer learning and discriminant correlation analysis, Inform. Fusion. 68 (2021) 131–148, doi:10.1016/j.inffus.2020.11.005.

[8] N.I.R. Yassin, S. Omran, E.M.F El Houby, H. Allam, Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: A systematic review, Comput. Method. Program. Biomed. 156 (2018) 25–45, doi:10.1016/j.cmpb.2017.12.012.

[9] W. Gomez, W.C.A. Pereira, A.F.C. Infantosi, Analysis of co-occurrence texture statistics as a function of gray-level quantization for classifying breast ultrasound, IEEE Trans. Med. Imaging. 31 (10) (2012) 1889–1899, doi:10.1109/TMI.2012.2206398.

[10] K.M. Amin, A.I. Shahin, Y. Guo, A novel breast tumor classification algorithm using neutrosophic score features, Measurement 81 (2016) 210–220, doi:10.1016/j.measurement.2015.12.013.

[11] T. Tan, B. Platel, H. Huisman, C.I. Sánchez, R. Mus, N. Karssemeijer, Computer-aided lesion diagnosis in automated 3-D breast ultrasound using coronal spiculation, IEEE Trans. Med. Imaging 31 (5) (2012) 1034–1042, doi:10.1109/TMI.2012.2184549.

[12] J. Bai, R. Posner, T. Wang, C. Yang, S. Nabavi, Applying deep learning in digital breast tomosynthesis for automatic breast cancer detection: a review, Med. Image Anal. 71 (2021) 102049, doi:10.1016/j.media.2021.102049.

[13] S.H. Wang, V.V. Govindaraj, J.M. Górriz, X. Zhang, Y.D. Zhang, Covid-19 classification by FGCNet with deep feature fusion from graph convolutional network and convolutional neural network, Inform. Fusion. 67 (2021) 208–229, doi:10.1016/j.inffus.2020.10.004.

[14] L. Bi, D.D. Feng, M. Fulham, J. Kim, Multi-Label classification of multi-modality skin lesion via hyper-connected convolutional neural network, Pattern Recogn. 107 (2020) 107502, doi:10.1016/j.patcog.2020.107502.

[15] Z. Xu, Y. Wang, M. Chen, Q. Zhang, Multi-region radiomics for artificially intelligent diagnosis of breast cancer using multimodal ultrasound, Comput. Biol. Med. 149 (2022) 105920, doi:10.1016/j.compbiomed.2022.105920.

[16] T. Xiao, L. Liu, K. Li, W. Qin, S. Yu, Z. Li, Comparison of transferred deep neural networks in ultrasonic breast masses discrimination, Biomed. Res. Int. 2018 (2018), doi:10.1155/2018/4605191.

[17] W.X. Liao, P. He, J. Hao, X.Y. Wang, R.L. Yang, D. An, L.G. Cui, Automatic identification of breast ultrasound image based on supervised block-based region segmentation algorithm and features combination migration deep learning model, IEEE J. Biomed. Health. 24 (4) (2019) 984–993, doi:10.1109/JBHI.2019.2960821.

[18] Z. Cao, L. Duan, G. Yang, T. Yue, Q. Chen, An experimental study on breast lesion detection and classification from ultrasound images using deep learning architectures, BMC Med. Imaging 19 (1) (2019) 51, doi:10.1186/s12880-019-0349-x.

[19] Z.M. Zhuang, Y.Q. Kang, A.N.J. Raj, Y. Yuan, W.L. Ding, S.M. Qiu, Breast ultrasound lesion classification based on image decomposition and transfer learning, Med. Phys. 47 (12) (2020) 6257–6269, doi:10.1002/mp.14510.

[20] M.I. Daoud, S. Abdel-Rahman, R. Alazrai, Breast ultrasound image classification using a pre-trained convolutional neural network, 15th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), 2019, doi:10.1109/SITIS.2019.00037.

[21] B. Zeimarani, M.G.F. Costa, N.Z. Nurani, S.R. Bianco, W.C.D. Pereira, C.F.F. Costa, Breast lesion classification in ultrasound images using deep convolutional neural network, IEEE Access 8 (2020) 133349–133359, doi:10.1109/ACCESS.2020.3010863.

[22] F. Wang, X. Liu, N. Yuan, B. Qian, L. Ruan, C. Yin, C. Jin, Study on automatic detection and classification of breast nodule using deep convolutional neural network system, J. Thoracic Dis. 12 (9) (2020) 4690–4701, doi:10.21037/jtd-19-3013.

[23] S. Han, H.K. Kang, J.Y. Jeong, et al., A deep learning framework for supporting the classification of breast lesions in ultrasound images, Phys. Med. Biol. 62 (19) (2017) 7714–7728, doi:10.1088/1361-6560/aa82ec.

[24] X. Yu, Q. Zhou, S. Wang, Y.D. Zhang, A systematic survey of deep learning in breast cancer, Int. J. Intell. Syst. 37 (1) (2022) 152–216, doi:10.1002/int.22622.

[25] Q. Huang, H. Luo, C. Yang, J. Li, Q. Deng, P. Liu, M. Fu, L. Li, X. Li, Anatomical prior based vertebra modelling for reappearance of human spines, Neurocomputing 500 (2022) 750–760, doi:10.1016/j.neucom.2022.05.033.

[26] Q. Huang, L. Ye, Multi-task/single-task joint learning of ultrasound BI-RADS features, IEEE T. Ultrason. Ferr. 69 (2) (2022) 691–701, doi:10.1109/TUFFC.2021.3132933.

[27] K. Wada, labelme: Image Polygonal Annotation with Python, 2016. https://github.com/wkentaro/labelme.

[28] T.Y. Lin, A. RoyChowdhury, S. Maji, Bilinear convolutional neural networks for fine-grained visual recognition, IEEE T. Pattern. Anal. 40 (6) (2017) 1309–1322, doi:10.1109/TPAMI.2017.2723400.

[29] F. Chollet, Xception: Deep learning with depthwise separable convolutions, in: 30th Proceedings of the IEEE conference on computer vision and pattern recognition, (CVPR), 2017, pp. 1251–1258, doi:10.1109/CVPR.2017.195.

[30] E. Lazarus, M.B. Mainiero, B. Schepps, S.L. Koelliker, L.S. Livingston, An automatic multi-view disease detection system via Collective Deep Region-based Feature Representation, Future Gener. Comp. S. Y. 115 (2021) 59–75, doi:10.1016/j.future.2020.08.038.

[31] Y.D. Zhang, Z. Dong, S.H. Wang, et al., Advances in multimodal data fusion in neuroimaging: overview, challenges, and novel orientation, Inform. Fusion. 64 (2020) 149–187, doi:10.1016/j.inffus.2020.07.006.

[32] A. Berger, S. Guda, Threshold optimization for F measure of macro-averaged precision and recall, Pattern Recognit. 102 (2020) 107250, doi:10.1016/j.patcog.2020.107250.

[33] A. Hijab, M.A. Rushdi, M.M. Gomaa, A. Eldeib, Breast cancer classification in ultrasound images using transfer learning, 2019 Fifth International Conference on Advances in Biomedical Engineering (ICABME), 2019, doi:10.1109/ICABME47164.2019.8940291.

[34] H. Tanaka, S. Chiu, T. Watanabe, S. Kaoku, T. Yamaguchi, Computer-aided diagnosis system for breast ultrasound images using deep learning, Phys. Med. Biol. 64 (23) (2019) 235013, doi:10.1088/1361-6560/ab5093.

[35] H. Kim, J. Park, H. Lee, G. Im, J. Lee, K.B. Lee, H.J. Lee, Classification for breast ultrasound using convolutional neural network with multiple time-domain feature maps, Appl. Sci. 11 (21) (2021) 10216, doi:10.3390/app112110216.

[36] Z. Zhuang, Z. Yang, A.N.J. Raj, C. Wei, P. Jin, S. Zhuang, Breast ultrasound tumor image classification using image decomposition and fusion based on adaptive multi-model spatial feature fusion, Comput. Method. Program. Biomed. 208 (2021) 106221, doi:10.1016/j.cmpb.2021.106221.

[37] Y. Luo, Q. Huang, X. Li, Segmentation information with attention integration for classification of breast tumor in ultrasound image, Pattern Recognit. 124 (2022) 108427, doi:10.1016/j.patcog.2021.108427.

[38] Q. Huang, Z. Miao, S. Zhou, C. Chang, X. Li, Dense prediction and local fusion of superpixels: a framework for breast anatomy segmentation in ultrasound image with scarce data, IEEE T. Instrum. Meas. 69 (1) (2022) 114–123, doi:10.1109/TIM.2021.3088421.

[39] J. Zhou, F. Pan, W. Li, H. Hu, W. Wang, Q. Huang, Feature fusion for diagnosis of atypical hepatocellular carcinoma in contrast-enhanced ultrasound, IEEE T. Ultrason. Ferr. 69 (1) (2022) 114–123, doi:10.1109/TUFFC.2021.3110590.

[40] Q. Huang, D. Wang, Z. Lu, S. Zhou, J. Li, L. Liu, C. Chang, A novel image-to-knowledge inference approach for automatically diagnosing tumors, Expert Syst. Appl. 229 (A) (2023) 120450, doi:10.1016/j.eswa.2023.120450.

**Yaozhong Luo** received the Ph.D. degree in information engineering from the School of Electronic and Information Engineering, South China University of Technology, in 2022. His research interests include ultrasonic image analysis, pattern recognition and machine learning.

**Qinghua Huang** received the Ph.D. degree in biomedical engineering from the Hong Kong Polytechnic University, Hong Kong, in 2007. In 2008, he joined the School of Electronic and Information Engineering, South China University of Technology, China. Now he is a full professor with School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include multi-dimensional ultrasonic imaging, pattern recognition, medical image analysis, machine learning for medical data, and intelligent computation.

**Longzhong Liu** is a full professor with Department of Ultrasound, The Cancer Center of Sun Yat-sen University, State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine, Guangzhou, Guangdong, China.