A Temporal Sequence Dual-Branch Network for Classifying Hybrid Ultrasound Data of Breast Cancer

ZIQI YANG, XUN GONG, YING GUO, AND WENBIN LIU

1School of Information Science and Technology, Southwest Jiaotong University, Chengdu 611756, China
2North China University of Science and Technology Affiliated Hospital, Tangshan 063000, China
3China Electronics Technologies Cyber Security Co., Ltd., Chengdu 610041, China

Corresponding author: Xun Gong (gongxun@foxmail.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61876158, and in part by the Sichuan Science and Technology Program under Grant 2019YFS0432.

ABSTRACT In clinical medicine, the contrast-enhanced ultrasound(CEUS) has been a commonly used imaging modality for diagnosis of breast tumor. However, most researchers in computer vision field only focus on B-mode ultrasound image which does not get good results. To improve the accuracy of classification, first, we propose a novel method, i.e., a Temporal Sequence Dual-Branch Network(TSDBN) which, for the first time, can use B-mode ultrasound data and CEUS data simultaneously. Second, we designed a new Gram matrix to model the temporal sequence, and then proposed a Temporal Sequence Regression Mechanism (TSRM), which is a novel method to extract the enhancement features from CEUS video based on the matrix. For B-mode ultrasound branch, we use the traditional ResNeXt network for feature extraction. While CEUS branch uses ResNeXt + R(2 + 1)D network as the backbone network. We propose a TSRM to learning temporal sequence relationship among frames, and design a Shuffle Temporal Sequence Mechanism(STSM) to shuffle temporal sequences, the purpose of which is to further enhance temporal information among frames. Experimental results show that the proposed TSRM could use temporal information effectively and the accuracy of TSDBN is higher than that of state-of-art approaches in breast cancer classification by nearly 4%.

INDEX TERMS Breast cancer classification, temporal sequence, contrast-enhanced ultrasound (CEUS), shuffle mechanism.

I. INTRODUCTION

Breast cancer is the most common cancer of women and the second leading cause of cancer death [1]. Early detection of breast cancer has been shown to significantly improve survival rate of patients [2], [3]. Therefore, correct diagnosis at early stage received widespread attention. Ultrasound has been widely used in the detection of early breast cancer because of its safety, low cost and high versatility [4]. However, its diagnostic accuracy depends on the special skills of the ultrasonic physicians—it says that the diagnosis difference could be larger than 30% among physicians of different levels [5].

In recent years, with the excellent performance of deep learning in image recognition, it has been widely used in ultrasound image classification and has achieved many progresses [6]–[11]. However, most data used by researchers is still B-mode ultrasound images. With the development of medical imaging, contrast-enhanced ultrasound (CEUS) videos can provide more precise pathological information by observing the dynamic enhancement of the lesion area in temporal sequences, and gradually becomes a more effective clinical diagnosis technology than traditional B-mode ultrasound, MRI and CT [12], [13]. Compared with B-mode ultrasound, the related research [14]–[16] show that the CEUS can visualize more sensitive imaging morphology and the flow of microvessels [17], hence, improving the classification accuracy between benign and malignant lesions. Obviously, CEUS contains enhanced information related to lesion that is helpful for breast cancer classification.

Fig. 1 is an example of our hybrid data, in (a) and (b), from left to right, each image is a frame of B-mode ultrasound...
To this end, we propose a novel method Temporal Sequence Dual-Branch Network (TSDBN), a network for breast cancer classification based on B-mode ultrasound video and CEUS video, the architecture of which is shown in Fig. 2. In the branch of B-mode ultrasound, we use the ResNeXt-18 [18] network to extract the morphological characteristics of breast lesions. In the branch of CEUS, to enhance the temporal feature of CEUS video, we design a Temporal Sequence Regression Mechanism (TSRM) and a Shuffle Temporal Sequence Mechanism (STSM), which make the network pay more attention to the discrepancy among frames along the timeline. First, the TSRM is proposed as a regression mechanism on temporal sequences that indicates the position of different frames in the video. The Gram matrix [19], which is widely used in the field of video generation, is used to express temporal sequences by calculating the distance between different frames in our TSRM block. At the same time, inspired by the method in the fine-grained image classification area [20], in order to enhance the temporal feature of the lesion area, a shuffle temporal sequence mechanism is proposed to disturb adjacent frames. Through STSM, the network will pay more attention to the critical information of CEUS that determine the temporal sequence, which is exactly the benefit that CEUS can provide.

The main contributions of this paper are as follows:

- To the best of our knowledge, for the first time, we proposed a dual-branch framework that uses hybrid data, i.e., B-mode ultrasound video and CEUS video, as input for breast cancer classification. Compared with state-of-art methods, our method has achieved the highest performance.
- A novel temporal feature extraction method, TSRM, of CEUS is proposed, which can extract the dynamic enhanced feature of the lesion area, and uses the shuffle temporal sequence to enhance the temporal feature of video.

This paper is organized by 5 sections: related work is analyzed in Section II. The proposed method is described in Section III. Experiments are conducted and discussed in Section IV. At last, the paper is concluded in Section V.

## II. RELATED WORKS

### A. BREAST CANCER CLASSIFICATION

Over recent decades, many researchers working on ultrasound have been trying to find a better solution to assist breast tumor diagnosis. Abdel-Nasser et al. [21] proposed the use of a super-resolution approach that exploit the complementary information provided by multiple images of the same target. The super-resolution-based approach improves the performance of the evaluated texture methods and thus outperforms the state of art in benign/malignant tumor classification. Alvarenga et al. [22] investigated seven morphological parameters in distinguishing malignant and benign breast tumors on ultrasound images and achieved a performance slightly over 83% in distinguishing malignant and benign breast tumors. Mohammed et al. [23] presented a fully computerized system (ANN based) to identify and discriminate the benign and malignant breast tumor cases by combining the ultrasound images and the experimental domain information of breast structure. Moreover, Gaussian process classifier is a powerful method for the direct uncertainty quantification of classification application. A breast cancer survivability prediction model that a hybrid of Incremental Learning radial basis function Neural Network, Gaussian Process classifier and AdaBoost can achieve higher prediction accuracy than conventional classifiers. Qi et al. [24] proposed a network to diagnose breast ultrasound images using deep convolutional
neural networks with multi-scale kernels and skip connections for improve sensitivity and robustness of classification. The network consists of two components to identify malignant tumors and recognize solid nodules in a cascade manner, which improve classification accuracy and sensitivity. Byra et al. [25] presented a matching layer for utilize a pre-trained model on the dataset with 3-channel natural images in grayscale ultrasound images. So, the aim of this layer is to rescale pixel intensities of the grayscale ultrasound images and convert those images to red, green, blue (RGB). An experiment results show the usefulness of the approach.

The main shortage of all those methods is that they were working on merely B-mode ultrasound images, lacking context information. Contrast-enhanced ultrasound (CEUS) is the application of ultrasound contrast medium to traditional medical sonography. CEUS has been proved to be more effective in early detection of tumor diagnosis in clinic applications [26]. In the field of ultrasound image analysis, the effectiveness of classification using CEUS data has been studied and proven [27]. Guo et al. [28] chosen three typical CEUS images from three phases of CEUS videos, which simulates the clinical diagnosis procedure of radiologists. Then, these images were fed to a multiple kernel learning (MKL) classifier. Pan et al. [29] directly used a 3D convolutional neural network (3D-CNN) to extract spatial and temporal features of CEUS. Meng et al. [30] presented a method of used B-mode ultrasound and CEUS to classification of liver tumor. Considering the specificity of the two data, the features are extracted from the B-mode ultrasound and CEUS separately, then the features is classified by a multiple empirical kernel learning machine (MEKLM) classifier, which can utilize information of the hybrid data. Although the method have made great achievements in aiding the diagnosis of liver cancer, the drawbacks are obvious. One is that the essential differences between CEUS and B-mode ultrasound have not been further studied. The second is that 3 images only selected from CEUS are not enough to represent the enhancement information of the lesion area. The third is that traditional machine learning methods are used to analyze this hybrid data. Based on this, we revisit many approaches to solve these problems and make further research. To the best of our knowledge, in the field of computer aided ultrasound diagnosis, CEUS video has not been used for automatic breast cancer classification. Therefore, for the first time, we use B-mode ultrasound and CEUS video simultaneously for breast cancer classification.

### B. TWO-STREAM METHOD

In the task of video classification based on two kinds of different data, the two-stream method is commonly used. For the first time, Simonyan and Zisserman [31] proposed a two-stream method which uses one stream to learn the spatial context of a single video frame and use another stream to model the motion characteristics from a stacked video optical flow. Then the average fusion is calculated from the softmax outputs of two branches. This method provides an instructive direction to combine multimodal data for classification. Further, Feichtenhofer et al. [32] analyzed the performance difference of the two-stream networks by using varying fusion strategies, like different ways of integrating spatial features and temporal features. Wang et al. [33] proposed a temporal segment network (TSN), which divides a long video into $n$ segments, then put $n$ segments into two streams respectively, and finally integrates the feature of $n$ segments for prediction. This approach aimed to solve the problem that long video is difficult to learn. Lan et al. [34] used the weights learned from TSN to evaluate the classification probability of different video segments. Zhou et al. [35] put forward a temporal relational network (TRN), which can learn the correlation of objects in the temporal domain between different frames.
through the network, so that the network is prone to recognize the primary actions. To combine different data for classification, the two-stream-based method can extract the feature of different data independently and fuse them properly. Inspired by the idea of two-stream method, we design a dual-branch network for our hybrid ultrasound data.

C. VIDEO UNDERSTANDING

In the last few years there has been great progress in the field of video understanding. For example, supervised learning and powerful deep learning models can be used to classify a number of possible actions in videos, summarizing the entire clip with a label. Feature representation is the core technique in video understanding. Besides the two-stream method, 3D convolution is another mainstream type of method. Inspired by the Inception-V1 [36], Carreira et al. [37] proposed I3D, where 3D convolution kernels of different sizes are used in each inception module and the $1 \times 1 \times 1$ convolution kernels were used for dimensionality reduction. Diba et al. [38] put forward the temporal 3D CNN(T3D) to solve the problem of insufficient information mining in the long time domain of 3D convolution. In the network, the author designed the Temporal Transition Layer(TTL) to replace the pooling layer, which has different temporal convolution kernel depths and can capture temporal feature-maps at different temporal depth ranges. Qiu et al. [39] proposed a Pseudo-3D ResNet(P3D ResNet), which uses a 2D space convolution of size $1 \times 1 \times 3$ and 1D time convolution of size $3 \times 1 \times 1$ instead of 3D convolution of size $3 \times 3 \times 3$, which can reduce the number of parameters and achieve better results. Based on the fact that the 2D convolution network has achieved the same accuracy as the 3D network in the field of motion recognition, Tran et al. [40] revisited the role of temporal reasoning in action recognition by means of 3D CNN, and proved that factorizing the 3D convolutional filters into separate spatial and temporal components yields significantly gains in accuracy. Finally, a new spatio-temporal convolutional block, R(2 + 1)D is designed, which produces CNN that achieve results superior to the P3D.

Compared with the previous networks are designed from the perspective of convolution along the timeline, some other networks are designed from the perspective of the particularity of video and have also achieved good results. Girdhar et al. [41] proposed an Action-VLAD pooling to replace the traditional average pooling and maximum pooling, which can aggregate evidence over the entire video about both the appearance of the scene and the motion of people without requiring every frame to be uniquely assigned to a single action. Considering that an action in most videos is independent of the background, Singh et al. [42] proposed a Multi-Stream Network(MSN), which uses a tracking algorithm to extract main object from the background. Along with the original image, the optical flow, the main object are input into a network of four branches. And then the Bi-directional LSTM network is used to extract the temporal feature of the images. As the motion of an object can be regarded as the graph structure of the spatio-temporal domain [43], Wang and Gupta [44] proposed the NGMN, which uses moving objects extracted from video frames to build graph structure, and then uses graph convolution to extract category information from the graph.

D. TEMPORAL SEQUENCE

As for CEUS, the fundamental difference from US is the temporal information provided. Video generation, which is a reversed problem of video analysis, can give us some hints to study temporal information. In order to generate coherent videos, a lot of research has been done on the temporal sequence. Hardy et al. [19] introduced the Gram matrix to model the dynamic transformation between consecutive frames, and used the Gram matrix as the motion feature to help network learn the dynamic between video frames. In order to adjust the relationship among frames in a time dimension, a temporal sequence association loss is designed [45], which is to ensure that there will not be too much discrepancy among frames of the video. To guarantee video coherence, the probabilities of start, middle and end points of the video sequence is modeled at the same time, to generate probabilities sequence of action start, action progress and action end [46]. Inspired by video generation, we design a CEUS branch in our network architecture, which uses a regression learning to mining the temporal sequence of CEUS.

III. THE PROPOSED METHOD

Clinically, the combination of B-mode ultrasound and CEUS has become a common technique for breast tumor diagnosis [47]. However, studies on both B-mode image and CEUS video are not well addressed in the field of computer aided ultrasound analysis, as it is hard to find a way to extract useful information from data of different modalities. This paper, a novel method Temporal Sequence Dual-Branch Network(TSDBN) is proposed to classify breast tumor by using both B-mode ultrasound and CEUS video, the architecture of which is shown in Fig. 2. The classical network ResNet-18 is used to extract image feature from B-mode ultrasound directly. For CEUS video, ResNet-18 + R(2 + 1)D [40] is taken as the backbone network. A Temporary Sequence Regression Mechanism(TSRM) and a Shuffle Temporal Sequence Mechanism(STSM) is proposed to promote the extraction capability from CEUS videos. Our network can effectively identify the difference between the original and the destructed CEUS videos, in this way, the temporal enhancement information can be further learned.

A. B-MODE ULTRASOUND AND CEUS DATA

In this paper, inspired by the uses of ultrasound in diagnostics [30], B-mode ultrasound and CEUS video are considered simultaneously to classify breast tumor. They are different expressions of the same lesion area and can help doctors get a better diagnostic image from more perspectives. B-mode ultrasound video riches in shape and texture, see Fig. 1(c),
but the pattern and brightness among adjacent frames are stable and rarely has variances. This characteristic of B-mode ultrasound means that there is no additional information in the time dimension. On the other hand, in the CEUS video, Fig. 1(d) has illustrated a clear pattern variances of among different frames in a short period, which means that the pattern in the temporal dimension is evident to provide more pathological information of the lesion area.

B-mode ultrasound image could provide the location, size, shape, internal echo, calcification, and other characteristics of the lesion area. CEUS video could provide dynamic status of the lesion area, including enhancement phase, enhancement intensity, enhancement sequence, enhancement lesion morphology, and other characteristics. Therefore, the B-mode ultrasound video only needs one frame to represent the whole video information. We choose a single frame with the maximum brightness value, denote as $S$. For CEUS video, in order to reduce the computational complexity and data redundancy, we need to select an appropriate number of frames to represent all the information of the original video as much as possible. Referring to the field of video understanding [35], [38], we use 16 as the number of extracted frames. The formula is as follows,

$$V_{ori} = \left\{ f^j_{bri} = \frac{[\max(f_{bri}) - \min(f_{bri})]}{16} \times i + f_{bri} \right\} \quad (1)$$

where $f_{bri}$ represents the brightness value of $j$-th frame, we first calculated the maximum($\max(f_{bri})$) and minimum($\min(f_{bri})$) value of brightness, then the corresponding frame is selected to from the set of frames($V_{ori}$) according to 16 equal division of brightness range. Finally, ($V_{ori}, S$) as an input to our network. In addition, $i \in \mathbb{N}$ and $0 < i < 16$.

Compared to natural image, lesion region has a rough boundary in B-mode ultrasound image and the contrast is low, which make it difficult to distinguish from the normal tissue. CEUS video is also different from general natural video, which does not contain any movements of an object, only the gradually enhancement of brightness and contrast affected by ultrasound contrast agents injected in the targeted tissues. So, the key is how to extract spatial features from B-mode ultrasound images and temporal features from CEUS video.

**B. OVERVIEW OF DUAL-BRANCH NETWORK**

As B-mode ultrasound and CEUS video are 2 different modalities, we should design one specific network for each type of data, and then combine them together as an end-to-end hybrid dual-branch network, which is capable of extracting the spatial and temporal features simultaneously.

In the branch of B-mode ultrasound, as shown in Fig. 2, ResNeXt-18 [18] is used as the texture and morphological feature extraction. The reason we choose ResNeXt-18 is that, at this stage, we only need to extract some basic and fundamental features, as the basic low-level morphological features are more useful in ultrasound classification. A very deep network will lead to too high-level features, which is not suitable for subsequent network to model temporal information. Moreover, ultrasound dataset is relatively small, a deep network will cause serious overfitting problem. In order to enhance the classification ability of the network, we concatenate the low-level and high-level features into a unified feature.

The shallow convolutional network can diminish the adverse effect of the jitter of CEUS video acquisition and the high noise characteristics of CEUS imaging by a shallow down-sampling. Therefore, in the CEUS branch, we also use ResNeXt-18 as the frame-level feature extractor for the reason. After all feature of 16 frames are obtained, which are then sent to the R($2 + 1$)D [40] to extract the temporal feature of this CEUS video. R($2 + 1$)D is a common and efficient method to extract temporal features. Compared with the $V_{ori}$, the frame feature obtained from ResNeXt-18 is more semantic and independent, and is more robustness for further exploiting temporal feature.

Then we concatenate the feature maps($f_{ori}$ and $f_{ce}$) extracted from $S$ and $V_{ori}$. After a convolution and a pooling layer, we got the probability vector of the corresponding category. The classification network loss function is defined as follows:

$$L_{cls} = - \sum_{V_{ori}, S} l \cdot \log[C(V_{ori}, S)] \quad (2)$$

where $F$ is the entire dataset, $C(V_{ori}, S)$ represents the classified network output of $V_{ori}$ and $S$ of the sample. $l = 0$ or 1, denotes the category labels, i.e., benign or malignant.

**C. TEMPORAL SEQUENCE REGRESSION MECHANISM**

When practitioner uses CEUS video to diagnose breast tumor, they mainly observe the enhancement process on images, along the timeline, of the lesion area, such as enhancement phase, enhancement intensity. The enhancement information of lesion areas is contained in different frames, and the different frames have sequence relationship in the time dimension. The sequence relationship is defined as temporal sequence. Therefore, the temporal sequence contains the enhancement information of the lesion area, and the corresponding temporal characteristics of the lesion area can be learned from the temporal sequence. Based on this, the Temporal Sequence Regression Mechanism(TSRM) proposed in CEUS branch to model sequence relationship among frames.

The core problem is to find a tool to express temporal sequences. In MD-GAN [19], Gram Matrix can be used to denote the correlation of two objects. Inspired by this idea, in this paper, the Gram Matrix is used to express the relationship of different frames. Another important key point is how to calculate the temporal correlation among frames. The temporal sequence correlation can be seen as the distance among frames in the time dimension, or discrepancy among frames. From this point of view, according to TGANS-C [45], a temporal sequence label is designed, as shown in Fig. 2(a). The distance between 2 frames is defined as follows:

$$D(f^i, f^j) = \text{Norm} \left( \|f^i - f^j\|_2^2 \right), \quad 0 < i, j < 16 \quad (3)$$
where $f^i$ and $f^j$ represent the $i$-th and $j$-th frames of a CEUS video, and then the L2-norm is used to measure the temporal sequence distance between 2 frames. The final label format is as follows.

$$M(V_{ori}) = \left[ \begin{array}{cccc} D(f^1, f^1) & D(f^2, f^1) & \cdots & D(f^{16}, f^1) \\ D(f^1, f^2) & D(f^2, f^2) & \cdots & D(f^{16}, f^2) \\ \vdots & \vdots & \ddots & \vdots \\ D(f^1, f^{16}) & D(f^2, f^{16}) & \cdots & D(f^{16}, f^{16}) \end{array} \right]$$ (4)

where $D(f^i, f^j)$ represents the distances of all pairs of frames, which can effectively express the enhancement information of the time dimension of video $V_{ori}$.

TSRM works on the feature map, and then an adaptive average pooling layer is utilized to get the $M(V_{ori})$ of size $16 \times 16$. And the TSRM loss is defined as:

$$L_{TSRM} = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} (G(V_{ori})(i,j) - M(V_{ori})(i,j))^2$$ (5)

where $0 < i, j < 16$. This loss calculates the difference between the predicted temporal sequence and the real sequence label. Through solving this regression problem, as we explained ahead, our CEUS branch will gain understandings of CEUS video, and pay more attention to the enhancing procedure of the lesion area in the video.

D. SHUFFLE TEMPORAL SEQUENCE MECHANISM

Shuffle mechanism is used in the field of natural language processing [48] and fine-grained image categorization [20], which local details play a more important role than global structures. The idea of shuffle mechanism could force the network to identify and focus on the discriminative local regions for recognition through destructing global structure and keeping local details. Similarly, if temporal sequence in a video are shuffled, discrepancy among frames that are critical to classification will enhance, and the network will be forced to classify video based on the discrepancy.

Therefore, the shuffle mechanism is used in our temporal sequence of $V_{ori}$. The principle of this mechanism is to deliberately reorder the 16 frames($f^i - f^{16}$) extracted from $V_{ori}$. However, destructing temporal sequence with STSM does not always bring beneficial information, which can lead the temporal sequence to be much confusion. With the use of TSRM, CEUS branch uses the temporal sequence label of $V_{des}$ for regression learning, hence, the network can understand the $V_{des}$ and learn the temporal information. There are two requirements for this mechanism. First, the temporal sequence should not be insufficient destructed, otherwise the $V_{des}$ and the $V_{ori}$ are uniform in temporal sequence information, which will lead to insufficient temporal information for network to learn. Second, the temporal sequence should not be over destructed, otherwise the discrepancy between temporal information of $V_{des}$ and the $V_{ori}$ is too large, in that case the network can not understand the temporal sequence information. Therefore, STSM only shuffles in the neighborhood of one frame, we have:

$$V_{des} = \left\{ f | \text{Shuffle} \left( \left\{ f^i, f^{i+1}, \cdots, f^{i+k} \right\} \right), f^i \in V_{ori} \right\}$$ (6)

where $V_{ori}$ represents the set of 16 frames selected from the CEUS video, $\text{Shuffle}()$ is a shuffle function used to shuffle the frames from $i$ to $i+k$ in $V_{ori}$, and the set of frames after STSM is $V_{des}$. In addition, $0 < i < 16 - k - 1$. By elaborately setting the value of $k$, we make sure that the shuffle is working only in the range of $k$ neighbors of current frame. It can effectively prevent over and insufficient destructed in $V_{ori}$. By shuffling the $V_{ori}$ properly, the network can not only focus on temporal information of the lesion area, but also solve the problem of data scarcity.

E. TOTAL LOSS

Our network has two outputs, one is classification probability, the other is a temporal sequence relationship matrix. The total loss is computed by:

$$L_{total} = \alpha L_{cls} + (1 - \alpha) L_{TSRM}$$ (7)

where $\alpha$ is designed to adjust the learning tendency of our network. By adjusting $\alpha$, the weight of $L_{cls}$ and $L_{TSRM}$ in total loss can be changed. Note that the TSRM and STSM block does not need to run in the prediction phase, this can greatly reduce the running time of the network when deploying a model.

IV. EXPERIMENTS

A. DATASET DESCRIPTION

Our hybrid ultrasound dataset consists of 268 samples, 146 are malignant and 122 are benign, each sample contains B-mode ultrasound video, CEUS video and pathological results. All data is collected from the ultrasound department of Sichuan province hospital in China. All samples are reliable and their labels, i.e., benign and malignant, were annotated by physicians. The paper divides the dataset into 10 subsets and uses 10-fold cross-validation to evaluate the performance of the proposed method.
Due to the particularity in the field of medical classification, the importance of each metric is not the same. e.g., $Rec$ weighs over others for tumor detection.

### C. PERFORMANCE COMPARISON

To assess the effectiveness of the proposed method, we design different comparison experiments. Since there is no literature on the breast cancer classification with CEUS, we choose the classical and the latest methods of video classification for comparison. All methods are implemented with the author publicly released open-source code, except TRN, LRCN, and NGMN, which code are not released online, we re-implement them in our experiments.

Results listed in Table. 1 has compared our methods with some state-of-art methods. It can be found that TSDBN_D has achieved the highest score in classification accuracy, which is 4% higher than other methods. At the same time, it has the highest score in a $Rec$, which can more effectively prevent the missed detection of breast tumor. And for $F1$, TSDBN_D also achieves the highest result, compared with the highest 90.2% of other methods, we increased by 3%.

In order to assess the role of CEUS video in different methods, three experiments are carried out: the first experiment only uses B-mode ultrasound image; the second only uses CEUS video; the third uses both data to classify breast tumor. From the results in Table. 1, from the 1st and 2nd row, the best $Acc$ is 82.6% using B-mode ultrasound, from 3rd and 4th row, the best result is 83.2% under CEUS video. Combining B-mode ultrasound image and CEUS video, our method can reach to the best $Acc$ of 90.2%. This is proved that the temporal information in CEUS video is helpful for breast cancer classification tasks, and the network proposed in the paper can effectively fuse the ultrasound image and CEUS video features together.

In the results of ablated models in Table. 1, we can find that the $Acc$ of the model decreases when STSM is added alone. It can be seen that the $V_{des}$ belongs to the wrong sample in the dataset. So the network can not extract the correct temporal information from the $V_{des}$, which leads to the decline of network accuracy. After adding TSRM, the $Acc$ of the model is improved by 2%, which shows that the temporal information extraction ability of our CEUS network can be effectively improved by the regression of learning temporal sequence. When STSM and TSRM are used together, the $Acc$ of the network is improved by 4% compared with the original model, and the final $Acc$ is up to 90.2%. $Rec$ and $Pre$ increased by 7.5% and 3.6% respectively, and the $F1$ increased by 4.8%. It can be seen that TSRM can learn the original temporal information of video from the $V_{des}$ by STSM.

The superiority of our method is illustrated more clear in Fig. 4, where (a) and (c) show ROC curves of our method and others. It can be seen that our method has achieved the highest results compared with others. Meanwhile, in the radar charts of (b) and (d), our method outperforms other methods in all four criterion. These results show that the

$$F1 = 2 \times \frac{Rec \times Pre}{Rec + Pre}$$  

(8)
method proposed in the paper is effective, and our method can learn useful temporal and spatial information from the hybrid data.

To more comprehensively measure our network performance, we compared TSDBN-D method and others in terms of parameters, model size, speed, accuracy. The speed is reported on one Nvidia GTX 1080Ti.

1) TEMPORAL FEATURE EXTRACTION NETWORK
Temporal feature extraction network is an important part of the CEUS branch. Different network have different feature extraction capabilities. In this paper, several classic temporal feature extraction networks are used, and the results are shown in Table 3. In this experiment, we keep the previous experimental settings unchanged, one difference is the temporal backbone network of CEUS branch. It can be seen from the table that R(2+1)D obtains the best result in our data. In addition, the methods based on 3D convolution are better than RNN can be found. After analysis, in video, to model temporal information and motion patterns of an object, RNN build temporal connections on the high-level features at the top layer while leaving the correlations in the low-level forms, e.g., edges at the bottom layers, not fully exploited. Compared with RNN, 3D convolution can perform temporal and spatial convolution directly on the frame to obtain more lower-level visual features for model temporal information. Specially, the CEUS video only contains the enhancement process of the lesion area but without motion information, which enhancement modeling is a low/mid-level operate that can be implemented via 3D convolutions. Therefore, 3D-based R(2+1)D is more suitable for CEUS video.

2) SHUFFLE GRANULARITY(K)
This is an important hyper parameter in our proposed method, which shows the extent of how we shuffle the temporal

D. MODEL ANALYSIS
The hyper parameters in the method have an impact on the results. These parameters are tunable and can directly affect how well a model can be trained. In this section, we will analyze all hyper parameters adopted in our method one by one.

VOLUME 8, 2020
82695
sequence. From Table 4, we can find that $K$ has a significant impact on classification accuracy. First, when $K$ value increases, our classification accuracy also increases. Begin from 1, $K$ keeps increasing, the classification accuracy begins to increase as well, and reaches the peak when $K = 3$. Generally speaking, if $K$ is too small, the discrepancies between the disturbed temporal sequence and the original temporal sequence are too small due to the similarity among frames. In that case, the network can not effectively learn the temporal information among different frames. On the contrary, if the $K$ is too large, the discrepancies between the disturbed temporal sequences will be too large, it is hard for the network to converge.

### TABLE 4. Performance of TSDBN with different $K$

| $K$ | Acc   | Rec   | Pre   | F1   |
|-----|-------|-------|-------|------|
| 1   | 83.3% | 87.4% | 82.4% | 84.7%|
| 2   | 86.0% | 87.0% | 94.3% | 90.2%|
| 3   | 90.2% | 91.4% | 95.2% | 93.2%|
| 4   | 87.7% | 83.2% | 92.4% | 83.5%|
| 5   | 84.2% | 80.1% | 91.2% | 81.2%|
| 7   | 82.3% | 79.0% | 89.3% | 79.1%|
| 8   | 79.0% | 75.0% | 85.0% | 76.2%|

### TABLE 5. The classification accuracy of the TSDBN trained with different proportion of $V_{ori}$ and $V_{des}$ in one batch.

| Ratio | Acc   | Rec   | Pre   | F1   |
|-------|-------|-------|-------|------|
| 1:0   | 86.2% | 83.7% | 91.9% | 88.4%|
| 1:1   | 90.2% | 91.4% | 95.2% | 93.2%|
| 1:2   | 87.3% | 87.6% | 93.5% | 88.2%|
| 1:3   | 86.6% | 85.2% | 91.2% | 87.6%|
| 0:1   | 84.3% | 84.4% | 92.4% | 84.7%|

3) **RATIO OF THE $V_{des}$ IN A MINI-BATCH**

$V_{des}$ is also a kind of unconventional data augmentation method, and its proportion in a min-batch also affects training results. The paper tests the classification accuracy under different proportions on CEUS videos. The results are shown in Table 5. When ratio of $V_{ori}$ and $V_{des}$ is set to 1:1 in a batch, the best results are obtained. Too much $V_{des}$ will reduce accuracy, which indicates that too high proportion of $V_{des}$ lead to too much chaos of temporal information. A ratio of 1:0 means STSM is not applied.

4) **IMAGE FEATURE EXTRACTION NETWORK**

In our method, image feature extraction network is an important part, which directly impacts the performance of the...
following temporal feature extraction. The classic VGG, ResNet and ResNeXt are chosen for comparison in this section, and the results are shown in Table 6. We find that, interestingly, higher performance can not be obtained by a deeper network, but a shallow network performs even better. Because only build temporal connections on the high-level features at the top layer while leaving the correlations in the low-level forms, e.g., edges at the bottom layers, not fully exploited. Therefore, the low-level features of the frame-level are more useful than high-level features in modeling CEUS videos. Namely, shallow network is more instrumental for our task. In addition, the low-level features of bottom layers can be transferred to the feature maps of top layers by the residual structure.

V. CONCLUSION
Medical ultrasound analysis has always been a challenging topic in computer vision and pattern recognition. The research in this field has been slow, due to the complexity of the ultrasound images and the lack of large ultrasound data. In this paper, To improve the accuracy of breast cancer classification by ultrasound, for the first time, we combine B-mode ultrasound and CEUS video together, which contain comprehensive and useful pathological information of the lesion area. For this hybrid data, a dual-branch network is proposed to extract spatial features from B-mode ultrasound video and temporal features from CEUS video. In the CEUS branch, we proposed TSRM based on temporal sequence in order to extract the pathological information of CEUS video more efficiently, which helps the network to concentrate on the region of lesion in the time dimension. Besides, inspired by the shuffle mechanism, the STSM is designed to enhance temporal information and data augmentation. Finally, the approach suggested in the paper produces the best results in our dataset.

Ultrasound images, like natural images, have uncertainties, which means that the same category may have different appearances, and the same appearance may be different categories. Therefore, to improve the classification ability, one is to improve the amount of train data, the other is to improve the learning ability of the network, including the identification of features and the robustness of the algorithm. In this paper, we mainly explore these two aspects, one is to increase the amount and types of data, the other is to design a network with powerful feature extraction ability.

Data is essential to train a good model for machine learning algorithms or neural networks. To make a better use of data, especially for medical images, it is necessary to design a method from the perspective of physicians. In medicine, it is found that the importance of CEUS video in physicians' pathological judgment is increasing. Therefore, in this work, we use CEUS to assist ultrasound in breast cancer classification, the results are especially promising. Our next work, hence, will still focus on exploiting useful information of CEUS via developing computer vision algorithms.

ACKNOWLEDGMENT
The authors would like to express their gratitude to all those who helped them during the writing of this article. They would like to thank Ms. Zhao and Mr. Irfan for their kind encouragement and useful instructions all through their writing.

REFERENCES

[1] R. L. Siegel, K. D. Miller, and A. Jemal, “Cancer statistics, 2016,” CA, A Cancer J. Clinicians, vol. 66, no. 1, pp. 7–30, 2016.
[2] J. Zhang, A. Saha, Z. Zhu, and M. A. Mazarowski, “Breast tumor segmentation in dce-mri using fully convolutional networks with an application in radiogenomics,” Proc. SPIE, vol. 10575, Feb. 2018, Art. no. 105750U.
[3] M. Yousefi, A. Krzyżak, and C. Y. Suen, “Mass detection in digital breast tomosynthesis data using convolutional neural networks and multiple instance learning,” Comput. Biol. Med., vol. 96, pp. 283–293, May 2018.
[4] M. Byra, T. Sznaider, D. Korzinek, H. Piotrzkowska-Wrobleswka, K. Dobruch-Sobczak, A. Nowicki, and K. Marasek, “Impact of ultrasound image reconstruction method on breast lesion classification with neural transfer learning,” 2018, arXiv:1804.02119. [Online]. Available: http://arxiv.org/abs/1804.02119
[5] R. J. Hooley, L. M. Scott, and L. E. Philpotts, “Breast ultrasonography: State of the art,” Radiology, vol. 268, no. 3, pp. 642–659, Sep. 2013.
[6] S. Blusuri, S. Jain, and J. Virmani, “Classification of breast lesions based on laws’ feature extraction techniques,” in Proc. 3rd Int. Conf. Comput. Sustain. Global Develop. (INDIACom), Mar. 2016, pp. 1700–1704.
[7] S. Han, H.-K. Kang, J.-Y. Jeong, M.-H. Park, W. Kim, W-C. Bang, and Y.-K. Seong, “A deep learning framework for supporting the classification of breast lesions in ultrasound images,” Phys. Med. Biol., vol. 62, no. 19, pp. 7714–7728, 2017.
[8] S. Y. Shin, S. Lee, I. D. Yun, S. M. Kim, and K. M. Lee, “Joint weakly and semi-supervised deep learning for localization and classification of masses in breast ultrasound images,” IEEE Trans. Med. Imag., vol. 38, no. 3, pp. 762–774, Mar. 2019.
[9] A. S. Becker, M. Mueller, E. Stoffel, M. Marcon, S. Ghafoor, and A. Boss, “Classification of breast cancer from ultrasound imaging using a generic deep learning analysis software: A pilot study,” Brit. J. Radiol., vol. 91, Dec. 2018, Art. no. 20170576.
[10] M. Byra, M. Galperin, H. Ojeda-Fournier, L. Olson, M. O’Boyle, C. Comstock, and M. P. Andre, “Comparison of deep learning and classical breast mass classification methods in ultrasound,” J. Acoust. Soc. Amer., vol. 146, no. 4, p. 2864, Oct. 2019.
[11] V. K. Singh, H. A. Rashwan, M. Abdel-Nasser, M. M. K. Sarker, F. J. Hooley, N. Pandher, S. Romani, and D. Puig, “An efficient solution for breast tumor segmentation and classification in ultrasound images using deep adversarial learning,” 2019, arXiv:1907.00887. [Online]. Available: http://arxiv.org/abs/1907.00887
[12] V. Cantisani, H. Grzhadhni, C. Fioravanti, M. Rosignuolo, F. Calliada, D. Messineo, M. G. Bernieri, A. Redler, C. Catalan, and F. D’Ambrosio, “Liver metastases: Contrast-enhanced ultrasound compared with computed tomography and magnetic resonance,” World J. Gastroenterol., vol. 20, no. 29, p. 998, 2014.

VOLUME 8, 2020

| Network    | Acc  | Rec  | Pre  | F1   |
|------------|------|------|------|------|
| VGG-16     | 85.7 % | 86.1 % | 85.8 % | 86.0 % |
| VGG-19     | 83.8 % | 84.2 % | 84.0 % | 83.9 % |
| ResNet-18  | 89.3 % | 89.4 % | 93.5 % | 91.7 % |
| ResNet-50  | 87.7 % | 88.0 % | 91.2 % | 90.0 % |
| ResNet-101 | 85.0 % | 85.5 % | 87.8 % | 88.7 % |
| ResNeXt-16(32×4d) | 90.2 % | 91.4 % | 95.2 % | 93.2 % |
| ResNeXt-50(32×4d) | 88.3 % | 88.9 % | 91.9 % | 90.4 % |
| ResNeXt-101(32×4d) | 85.7 % | 85.9 % | 88.1 % | 89.2 % |
F. Meng, J. Shi, B. Gong, Q. Zhang, L. Guo, D. Wang, and H. Xu, “B-mode ultrasound based diagnosis of liver cancer with CEUS images as privileged information,” in Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC), Jul. 2017, pp. 3117–3121.

S. Balocco, C. Gatta, O. Pujol, J. Mauri, and P. Radeva, “SRBF: Speckle reducing bilateral filtering,” Ultrasound Med. Biol., vol. 39, no. 3, pp. 355–368, 2013.

S. M. Pizer, E. P. Amburn, J. D. Austin, R. Cromartie, A. Geselowitz, B. ter Haar Romeny, J. B. Zimmerman, K. Zuiderveld, “Adaptive histogram equalization and its variations,” IEEE Trans. Med. Imag., vol. 13, no. 3, pp. 297–311, Mar. 2006.

T. Greer, B. ter Haar Romeny, J. B. Zimmerman, K. Zuiderveld, “Adaptive reducing bilateral filtering,” IEEE Trans. Med. Imag., vol. 24, no. 11, pp. 1309–1320, Nov. 2005.

D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, “A closer look at spatiotemporal convolutions for action recognition,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jul. 2017, pp. 971–980.

B. Singh, T. K. Marks, M. Jones, O. Tuzel, and M. Shao, “A multi-stream bi-directional recurrent neural network for fine-grained action detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 6450–6459.

R. Girdhar, D. Ramanan, A. Gupta, J. Sivic, and B. Russell, “Action-VR: Learning spatio-temporal aggregation for action classification,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 971–980.

S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, “Aggregated residual transformations for deep neural networks,” 2016, arXiv:1611.05431. [Online]. Available: http://arxiv.org/abs/1611.05431

M. Byra, M. Galperin, H. Ojeda-Fournier, L. Olson, M. O’Boyle, M. Wubulihasimu, M. Maimaitusun, X.-L. Xu, X.-D. Liu, and Y. Miyamoto, T. Ito, E. Takada, K. Omoto, T. Hirai, and F. Moriyasu, “Going deeper with convolutions,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 1–9.

L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. van Gool, “Temporal segment networks: Towards good practices for deep action recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 1979–1987.

J. Shi, B. Gong, Q. Zhang, L. Guo, D. Wang, and H. Xu, “B-mode ultrasound based diagnosis of liver cancer with CEUS images as privileged information,” in Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC), Jul. 2019, pp. 1359–1363, Aug. 2010.

S. Balocco, C. Gatta, O. Pujol, J. Mauri, and P. Radeva, “SRBF: Speckle reducing bilateral filtering,” Ultrasound Med. Biol., vol. 36, no. 8, pp. 1353–1363, Aug. 2010.

Z. Yang et al.: TSDBN for Classifying Hybrid Ultrasound Data of Breast Cancer

IEEE Access VOLUME 8, 2020

82698
[54] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning spatiotemporal features with 3D convolutional networks,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 4489–4497.
[55] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural Comput., vol. 9, no. 8, pp. 1735–1780, 1997.
[56] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, “Empirical evaluation of gated recurrent neural networks on sequence modeling,” 2014, arXiv:1412.3555. [Online]. Available: http://arxiv.org/abs/1412.3555

ZIQI YANG received the B.S. degree in computer science from the Weifang University of Science and Technology. He is currently pursuing the M.S. degree with the Department of Computer Science, Southwest Jiaotong University. His research interests include pattern recognition, computer vision, and medical image processing.

XUN GONG received the B.S. degree in computer science and technology from Beijing Technology and Business University, in 2003, and the Ph.D. degree in computer science and technology from Southwest Jiaotong University (SWJTU), China, in 2008. He was a Visiting Scholar with Alberta University, Canada, in 2015, Louisiana State University, USA, from July 2018 to February 2019. He is currently an Associate Professor with the School of Information Science and Technology, Southwest Jiaotong University. His research interests include pattern recognition, computer vision, medical image processing, and deep learning.

WENBIN LIU received the B.S. degree in communication engineering from Southwest Jiaotong University, in 2005, and the master’s degree in communication and information system from the Beijing University of Posts and Telecommunications, in 2008. He is currently pursuing the Ph.D. degree with the School of Information Science and Technology, Southwest Jiaotong University. He is currently working as a Senior Engineer with China Electronics Technology Cyber Security Company Ltd. His research interests include information security, signal processing, and deep learning.

YING GUO received the master’s degree from Jinzhou Medical University. She is currently an Ultrasound Doctor with the North China University of Science and Technology Affiliated Hospital. Her research interests include image diagnosis and research of heart disease, thyroid disease, and breast disease.