Domain Knowledge Powered Deep Learning for Breast Cancer Diagnosis Based on Contrast-Enhanced Ultrasound Videos

Chen Chen, Yong Wang, Jianwei Niu, *Senior Member, IEEE*, Xuefeng Liu, Qingfeng Li, and Xuantong Gong

***Abstract*— In recent years, deep learning has been widely used in breast cancer diagnosis, and many high-performance models have emerged. However, most of the existing deep learning models are mainly based on static breast ultrasound (US) images. In actual diagnostic process, contrast-enhanced ultrasound (CEUS) is a commonly used technique by radiologists. Compared with static breast US images, CEUS videos can provide more detailed blood supply information of tumors, and therefore can help radiologists make a more accurate diagnosis. In this paper, we propose a novel diagnosis model based on CEUS videos. The backbone of the model is a 3D convolutional neural network. More specifically, we notice that radiologists generally follow two specific patterns when browsing CEUS videos. One pattern is that they focus on specific time slots, and the other is that they pay attention to the differences between the CEUS frames and the corresponding US images. To incorporate these two patterns into our deep learning model, we design a domain-knowledge-guided temporal attention module and a channel attention module. We validate our model on our Breast-CEUS dataset composed of 221 cases. The result shows that our model can achieve a sensitivity of 97.2% and an accuracy of 86.3%. In particular, the incorporation of domain knowledge leads to a 3.5% improvement in sensitivity and a 6.0% improvement in specificity. Finally, we also prove the validity of two domain knowledge modules in the 3D convolutional neural network (C3D) and the 3D ResNet (R3D).**

***Index Terms*— 3D convolution, attention mechanism, breast cancer, contrast-enhanced ultrasound, domain knowledge.**

Manuscript received April 14, 2021; revised April 22, 2021; accepted April 28, 2021. Date of publication May 7, 2021; date of current version August 31, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 61976012 and Grant 61772060, in part by the China Postdoctoral Science Foundation under Grant 2017M620683, and in part by the Beijing Hope Run Special Fund of Cancer Foundation of China under Grant LC2019A01. *(Yong Wang is co-first author.) (Corresponding author: Xuefeng Liu.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Independent Ethics Committee of National Cancer Center/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College under Application No. NCC2019C-178.

Chen Chen, Jianwei Niu, and Xuefeng Liu are with the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing 100191, China, and also with the Beijing Advanced Innovation Center for Big Data and Brain Computing (BDBC), Beihang University, Beijing 100191, China (e-mail: zy1906702@buaa.edu.cn; niujianwei@buaa.edu.cn; liu_xuefeng@buaa.edu.cn).

Yong Wong and Xuantong Gong are with the Department of Ultrasound, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100021, China (e-mail: drwangyong77@163.com; gongxuantong@163.com).

Qingfeng Li is with the Research Center of Big Data and Computational Intelligence, Hangzhou Innovation Institute, Beihang University, Hangzhou 310051, China (e-mail: liqingfeng@buaa.edu.cn).

Digital Object Identifier 10.1109/TMI.2021.3078370

I. Introduction

DEEP learning has been widely applied to computer-aided diagnosis (CAD), and these CAD systems based on deep learning have achieved extensive success in a variety of subfields such as the diagnosis of liver cancer [1], breast cancer [2], and cataract [3]. Especially for breast cancer diagnosis, there are numerous models that have achieved superior performance, such as ME-CNN [4], SaNet [5] and COAM [6]. Most of the models are based on static ultrasound (US) images due to the non-invasive and inexpensive nature of breast US examinations [7], [8].

However, in practice, besides the static breast US images mentioned above, radiologists would consider multimodality data such as elastography [9], [10] or contrast-enhanced ultrasound (CEUS) for a more accurate diagnosis [11]. The elastography technique provides information on tissue biomarkers such as elasticity, viscosity, and tissue nonlinearity. And the CEUS technique can reveal detailed information on tumors' blood supply [12], [13], compared with static US images. When performing CEUS examination, radiologists perform US examination at the same time to observe tumor changes by comparison. During CEUS examinations, the contrast agent is first injected into a tumor area, and then radiologists observe brightness changes in the tumor area through the CEUS video, as shown in Fig. 1 (a) [12], [14]. As the brightness change of CEUS frames reveals the blood supply of a tumor, the CEUS video contains important information for classifying tumors as benign or malignant. Despite the popularity of CEUS, to the best of our knowledge, there are few deep learning models based on CEUS videos for breast cancer diagnosis.

A straightforward approach to design a deep learning model based on CEUS videos is to adopt the state-of-the-art models initially designed for video classification tasks (e.g., ir-CSN [15], HAF + BoW [16]). More specifically, many of the models above utilize temporal attention modules to help

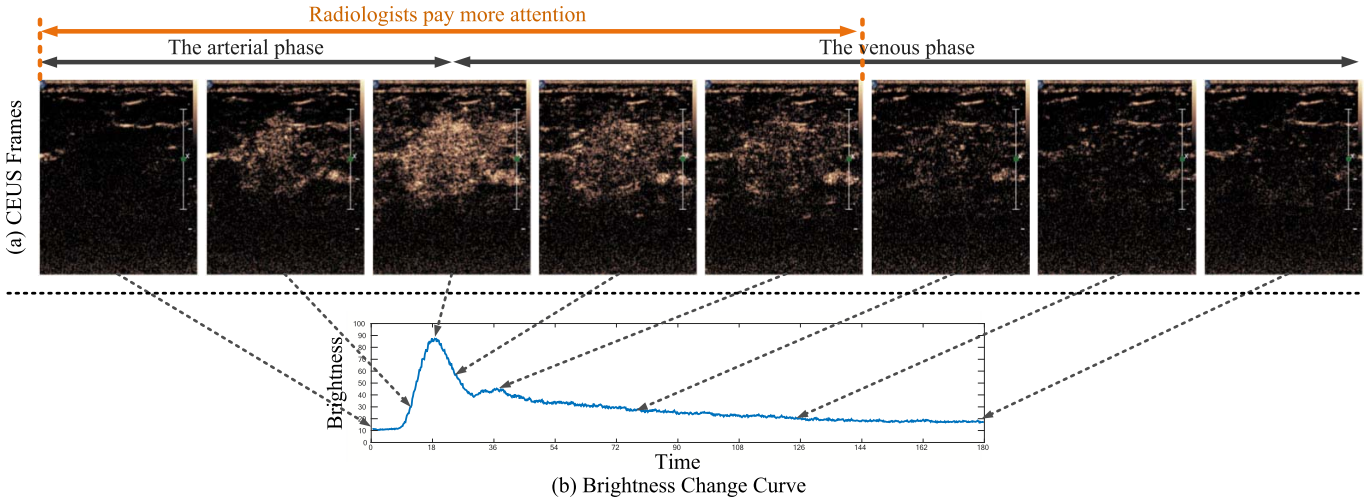


Fig. 1. The brightness change in a CUES video. (a) A few CEUS frames of one CEUS video. The front part is the frames of the injection, and the back part is the frames of washing out. (b) The brightness change curve of the breast CEUS video. In this figure, we can see that the brightness changes significantly during the initial arterial phase and the first half of the venous phase.

models focus on important time slots of videos to improve their performance [17], [18]. However, these models will encounter difficulty if they are applied to analyze CEUS videos. In particular, they require large-scale datasets, such as Kinetics-600 [19] and UCF-101 [20]. Unlike the large-scale datasets for video classification tasks, CEUS video datasets are generally much smaller because of the high cost and the difficulty of collecting the training data, and the extensive storage usage required. For example, a 3-minute CEUS video usually takes 400-500MB. We construct a CEUS dataset called Breast-CEUS which contains 221 CEUS videos from 217 patients and occupies about 120GB of storage space. This is already the largest breast CEUS dataset we could find.

To address the problem of small-scale training data, we integrate domain knowledge of radiologists into the deep learning model. Through discussions with our collaborators in the Cancer Hospital, we find that radiologists generally follow some specific patterns when browsing CEUS videos.

One of the most important patterns is that radiologists always focus on some specific time slots in CEUS videos. As shown in Fig. 1 (b), a CEUS video can be divided into the arterial phase and the venous phase. In the arterial phase, the contrast agent enters the tumor gradually, and in the venous phase it washes out gradually. Radiologists generally pay more attention to the arterial phase and the first half of the venous phase. This is because most of the blood supply information of the tumors is contained in these time slots. In conjunction with Fig. 1 (b), we can see that radiologists' attention is roughly consistent with the brightness of the video, so the brightness can be utilized to represent radiologists' attention.

It is interesting to test whether the model can learn radiologists' attention after training on the CEUS dataset. We adopt a deep learning model with a temporal attention module (TAM) which is similar to Zhu *et al.* [21]. The TAM calculates temporal attention for each time slot in the input feature map. The temporal attention weight of a time slot directly indicates how much attention the model pays to the videos in this time

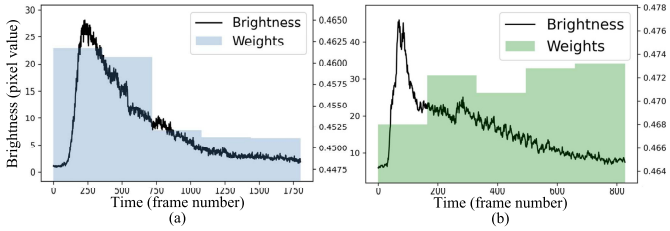


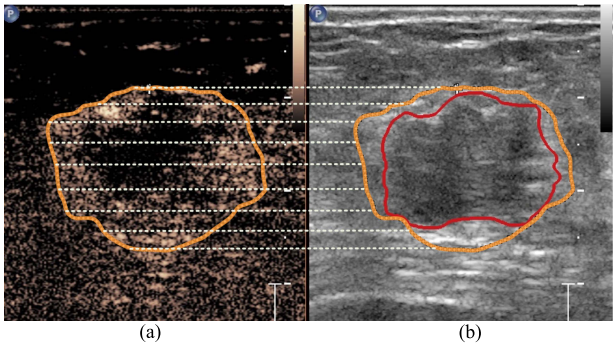
Fig. 2. The match between temporal attention and brightness change curve of two cases. (a) shows a correctly classified case and (b) shows a misclassified one. The temporal attention is shown as the histogram, and the brightness change is shown as the curve. The temporal attention and brightness in the correctly classified case match better than that in the misclassified case.

slot. Fig. 2 illustrates the attention of the model (represented as the temporal attention weight in the TAM), and that of radiologists (represented as the brightness of CEUS videos) of the two specific cases. The case on the left is correctly classified, and we can see that the attention of the model matches well with that of radiologists. The case on the right is misclassified and the attention of the model and radiologists' do not match well. It should be noted that whether the temporal attention of the model matches that of the radiologist is determined by the radiologist manually. The statistics for the dataset are shown in Table I. For those correctly classified cases, the temporal attention of 40.4% cases matches with radiologists' experience. In contrast, for those misclassified cases, the temporal attention of 79.0% cases does not match with radiologists' experience. In addition, we perform a correlation analysis using a chi-square test for classification results and the match between the temporal attention of the model and radiologists. The results indicate that the correlation is statistically significant with 95% confidence intervals. Therefore, we believe the accuracy can be improved if the model can focus on the time slots that radiologists pay attention to.

Another specific pattern is that radiologists observe CEUS videos to see whether the tumor area in CEUS frames is

TABLE I
THE CONTINGENCY TABLE SHOWS THE CORRELATION BETWEEN CLASSIFICATION RESULTS AND TEMPORAL ATTENTION MATCHING

Classification results	Whether the temporal attention of the model matches that of the radiologist		Total
	No	Yes	
Wrong	49	13	62
Correct	124	84	208
Total	173	97	270



(a) (b)

Fig. 3. The enlarged boundary of the malignant tumor between the CEUS window and the US window. (a) The boundary is drawn by radiologists based on CEUS. (b) The inner boundary is drawn by radiologists based on US. The outer boundary corresponds to the boundary in (a). The enlarged area is caused by the infiltration of a malignant tumor.

larger than that in the corresponding US images, as shown in Fig. 3. Due to the infiltrative nature of malignant tumors, the risk of tumors being malignant would be increased if the tumor diameter in CEUS window is greater than that in US window [22]–[25].

To integrate these two diagnostic patterns into our deep learning model, we investigate studies on multimodal fusion [26]–[28]. Since the method of concatenating ignores the correlation between multimodal features and the method of bilinear pooling outputs features with huge dimensions, we choose the attention mechanism and design two attention modules. In particular, we design a domain-knowledge-guided temporal attention module (DKG-TAM) to keep the model focusing more on the arterial and the first half of the venous phase, as radiologists do during diagnosis. In addition, we extract the difference in the diameter of the tumor shown respectively in the US window and the CEUS window. Differences in some first-order features and other features in two windows are also extracted, such as energy, perimeter, and variance. Then we design a domain-knowledge-guided channel attention module (DKG-CAM) to incorporate the extracted features into the model.

We validate our model on Breast-CEUS dataset composed of 221 cases. The result shows that utilizing the DKG-TAM and DKG-CAM simultaneously can achieve a sensitivity of 97.2% and an accuracy of 86.3%, which is a 3.5% improvement and a 6.0% improvement compared with our vanilla network. While adding DKG-TAM or DKG-CAM alone can outperform the vanilla network by 2.1% or 2.6% in accuracy, respectively.

We summarize the contributions of this paper as follows:

(i) We find two specific diagnostic patterns radiologists generally follow in temporal and feature.
(ii) Based on two diagnostic patterns of radiologists, we design two modules (DKG-TAM and DKG-CAM) to incorporate domain knowledge into a deep learning model.
(iii) We validate our model on Breast-CEUS dataset, and the result demonstrates that radiologists' domain knowledge can help in improving the performance of neural networks.

II. RELATED WORKS

A. CAD in Breast Cancer

In recent years, with the development of computer technology, deep learning techniques have been widely used in CAD systems for breast cancer [4], [29]–[39]. These CAD systems can be divided into five categories according to the modalities of data, mammogram (MG) [31]–[33], [40], ultrasound (US) [30], [34], [35], magnetic resonance imaging (MRI) [4], histopathologic (HG) images [36], [37], and multi-modalities [38], [39]. Some of these systems mentioned above have achieved excellent results, such as ME-CNN [4], MSGRAP [34], and DL-CNN-FCRN [31]. In particular, ME-CNN is a mixture ensemble of convolutional neural network which achieves an accuracy of 96.4% and a sensitivity of 97.7% on breast MRI images. MSGRAP is an architecture including a channel attention module with multi-scale grid average pooling. In the breast cancer segmentation, MSGRAP achieves an accuracy of 97.8% and a sensitivity of 80.4% on US images. DL-CNN-FCRN is a deep learning convolutional neural network combined with fully complex-valued relaxation network. DL-CNN-FCRN achieves an accuracy of 99% and a sensitivity of 98.8% on mammogram images.

Although these systems use different modalities of data, these data are essentially static images. In the actual diagnosis of breast cancer, radiologists not only perform US examinations but also perform CEUS examinations. Unlike the US examination, CEUS records a 3-minute video that reflects the blood supply of the tumor area. By analyzing the CEUS videos, the radiologists could make an accurate diagnosis. Although there are many CAD systems that use breast images for classification diagnosis, there are few relevant studies using CEUS videos.

B. CAD Based on Breast CEUS

CEUS is a novel technique based on US, reflecting the shape characteristics of the tumor and facilitating the observation of the tumor's blood supply. Owing to the high vascular density and disorganized distribution of breast tumors, CEUS, which shows the blood supply information, can provide valuable information for the diagnosis of breast tumors [12], [24], [41].

CEUS initially is used for liver tumor examinations but has been widely applied to breast tumor examinations in recent years [42]. There have been a few kinds of research combining breast CEUS with CAD. The work closest to our

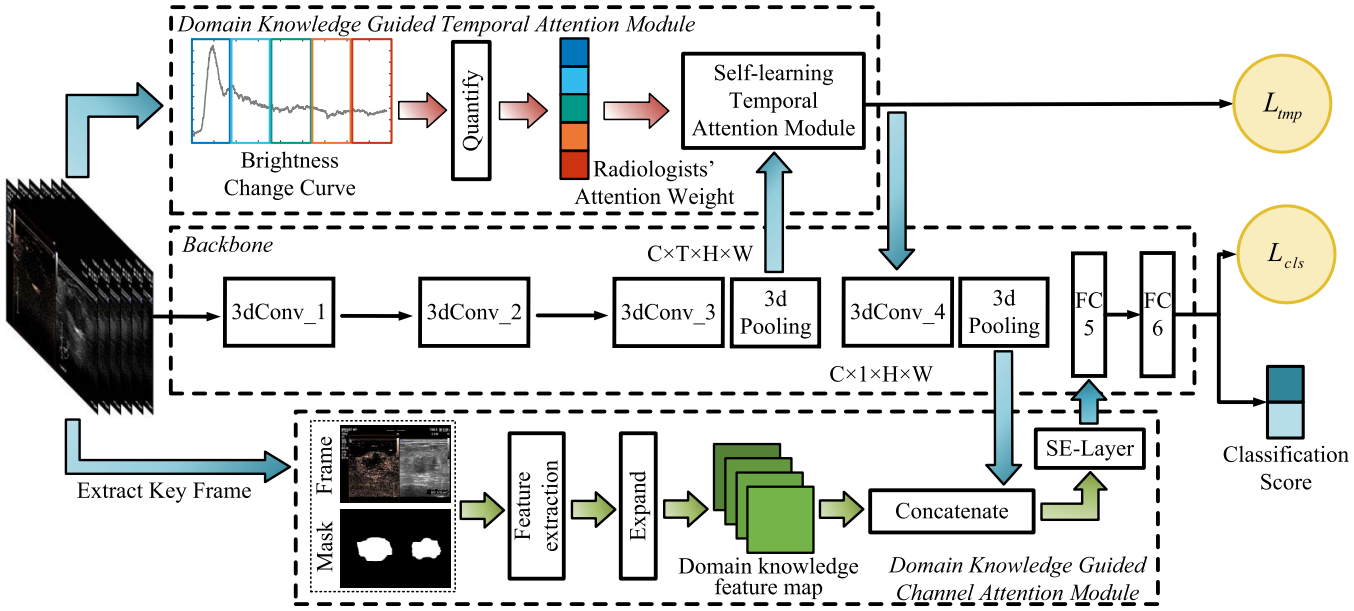


Fig. 4. The architecture of the model. It consists of three parts, including Backbone, DKG-TAM, and DKG-CAM. The Backbone is a variant of C3D. The DKG-TAM uses L_{tmp} and domain knowledge to constrain temporal attention learned by the model. The DKG-CAM expends the number of features for classification and reallocates the weights corresponding to each feature.

research work is that of Yang *et al.* [43]. They proposed a dual-branch network to extract spatial features from B-mode ultrasound images and designed a Gram matrix to model the temporal sequence from CEUS videos. But they did not consider incorporating the radiologists' domain knowledge into the model. In this paper, we experimentally demonstrate that the radiologists' domain knowledge can improve neural network's performance.

III. Method

In this section, we first present the overview of our breast cancer classification model, followed by a step-by-step presentation of the modules of our model and the detailed structure.

A. Overview

The overall model is shown in Fig. 4. The model is divided into three modules: the backbone 3D convolutional neural network (C3D), a domain-knowledge-guided temporal attention module (DKG-TAM), and a domain-knowledge-guided channel attention module (DKG-CAM). The backbone extract temporal and spatial features from the input CEUS videos. DKG-TAM integrates the radiologists' temporal attention into the model, guiding the model to focus on critical time slots in the CEUS videos. DKG-CAM is used to concatenate the features extracted based on the radiologist's domain knowledge with the feature extracted by the C3D backbone. Then the SE-Layer [44] in CAM is used to allocate weights to the different features in the dimension of the channel. The detailed structure of each module is described in the following sections.

B. Vanilla C3D Backbone

Since both C3D and R3D models are overfitting the training data, we reduce the number of layers in C3D and construct our own backbone from C3D. The backbone can automatically learn both temporal and spatial features from the video simultaneously. The structure of the vanilla C3D backbone we built is shown as *Backbone* in Fig. 4.

The architecture of the C3D backbone includes four 3D convolution layers, two pooling layers, four batch-normalized layers, two fully-connected layers, five dropout layers, and a softmax layer. All the convolutional layers are with stride size of $2 \times 2 \times 2$, and each convolutional layer is followed by a batch-normalized layer and a dropout layer. In addition, the third 3D convolution layer and the fourth 3D convolution layer both have a pooling layer behind them. Among them, the first 3D convolutional layer has a kernel size of $5 \times 5 \times 5$, and the remaining three 3D convolutional layers have a kernel size of $2 \times 2 \times 2$. The kernel size of pooling layers is $2 \times 2 \times 2$. Besides, the ratio of the dropout layer is fixed at 0.1.

In this paper, the input of our C3D backbone is $240 \times 320 \times 100$, which means that an input video is equally spaced to get 100 video frames, and each frame is resized to 240×320. After these convolutions and poolings, the input data are converted into a $5 \times 3 \times 2$ feature map. We expand the feature map into a 288-dimensional vector and feed it into the full connectivity layer (FC5) to get a 256-dimensional vector. Between FC5 and the next fully connected layer, there is a dropout layer to avoid over fitting. Eventually, the fully connected layer (FC6) outputs category scores.

C. Domain-Knowledge-Guided Temporal Attention Module

Since each frame is regarded equally important in C3D backbone, we propose a temporal attention module (TAM) to pay more attention to the key frames in the video. As shown in Fig. 4, the TAM is added to the end of the third

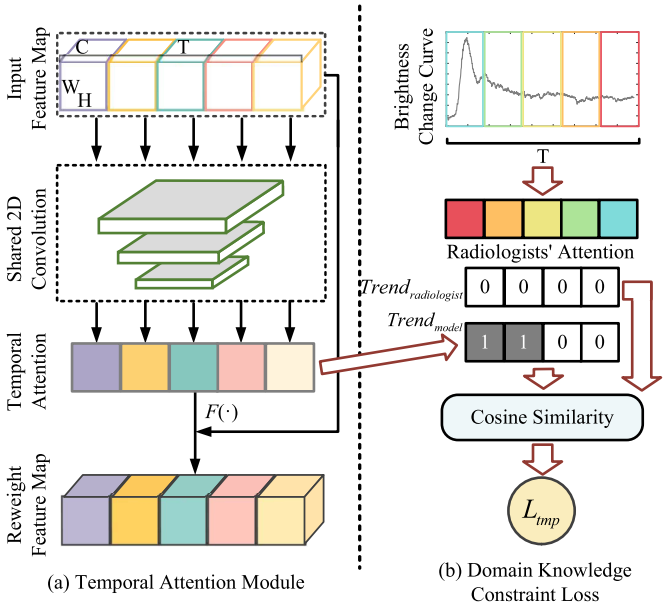


Fig. 5. The structure of DKG-TAM. It includes temporal attention module and domain knowledge constraint loss. (a) The structure of TAM. It uses feature maps and shared convolution layers to calculate the temporal attention of the model. (b) The domain knowledge constraint loss. It is calculated by the cosine similarity between $trend_{radiologist}$ and $trend_{model}$ and performing a negative logarithm transform.

3D convolutional layer. The TAM contains three successive 2D convolutional layers and a temporal-wise multiplication operation, as shown in Fig. 5 (a). The first 2D convolutional layer has a kernel size of 3×3 with stride 2×2 and outputs a 24 channels feature map. The second 2D convolutional layer has a kernel size of 2×2 with stride 2×2 and outputs a feature map with 8 channels. The last 2D convolutional layer has a kernel size of 3×4 with stride 1×1 and outputs the temporal weights.

The input feature map of size C×T×H×W outputs temporal weights of size $1 \times 1 \times$ T after passing through these three convolutional layers. The learned temporal attention is multiplied in the temporal dimension with the input feature map, and the final output is a feature map that has undergone temporal attention weighting and is consistent with the size of the input feature map.

However, we notice that there are still some misclassified cases, in which TAM does not learn the radiologists' temporal attention well, as shown in Fig. 2 (b). On the basis of TAM, we propose a domain-knowledge-guided temporal attention module (DKG-TAM) with domain knowledge to constrain temporal attention. We quantify radiologists' temporal attention from the brightness change curve, calculate the trend of corresponding temporal attention change, and design a loss function to constrain the TAM learn radiologists' temporal attention change trend, as shown in Fig. 5 (b). We refer to the TAM guided by domain knowledge as DKG-TAM.

The details of the process are as follows. Since the change in radiologists' attention during video viewing is related to the degree of contrast agent entry, which is directly related to the brightness magnitude in the video. Therefore, we choose to quantify radiologists' attention from the brightness change curve of the video. The brightness change is the average of the total images instead of the average of some specific pixels of the lesion. The module divides the brightness change curve of the raw input video into T (the input temporal dimension) windows at equal intervals. Then we downscale the brightness change within each window, such as calculating the mean or range of the brightness change for each time slot and obtain a $1 \times 1 \times$ T vector of radiologists' attention.

TABLE II
THE DOMAIN KNOWLEDGE FEATURES USED BY DKG-CAM

Type	Subtype
First-order	Energy, Entropy, Interquartile Range, Kurtosis, Maximum, Mean, Mean Absolute Deviation, Median, Range, Robust Mean Absolute Deviation, Root Mean Squared, Skewness, Total Energy, Uniformity, Variance
Shape	Elongation, Major Axis Length, Maximum Diameter, Mesh Surface, Minor Axis Length, Perimeter, Perimeter Surface Ratio, Pixel Surface, Sphericity,
GLSZM	Large Area Emphasis, Large Area High Gray Level Emphasis, Zone Entropy, Zone Variance
GLDM	Dependence Non Uniformity, Gray Level Non Uniformity
GLRLM	Run Entropy
NGTDM	Busyness

$$trend_i = \begin{cases} 0 & attention_i \geq attention_{i+1} \\ 1 & attention_i < attention_{i+1} \end{cases} \quad (1)$$

Since temporal attention only reflect the trend of radiologists' attention allocation, it is difficult to give a specific numerical magnitude of attention allocation. Therefore directly applying the radiologists' attention weights to the feature maps is inappropriate. Considering this factor, we calculate the attention change trend vector $trend_{radiologist}$ and $trend_{model}$ for radiologists and model by the Eq. 1, where $trend_i = 1$ indicates rising from time slot i to time slot $i+1$ and $trend_i = 0$ indicates falling. L_{tmp} is calculated by the cosine similarity between $trend_{radiologist}$ and $trend_{model}$ and performing a negative logarithm transform, as shown in Eq. 2. $trend_{model}$ and $trend_{radiologist}$ indicate the temporal change trend of the radiologists and the model, respectively. Finally, we add the loss of temporal attention (L_{tmp}) with the loss of classification (L_{cls}, Eq. 3) to obtain the overall loss (L), as shown in Eq. 4:

$$L_{cls}(x_n, y_n) = -w_n \{y_n \cdot \log \sigma(x_n) + (1 - y_n) \cdot \log(1 - \sigma(x_n))\} \quad (3)$$

$$L = L_{tmp} + L_{cls} \quad (4)$$

L_{cls} is a binary cross entropy loss, where n is the batch size, σ refers to the softmax [45] function, w_n denotes the weights of this batch, y_n is the target of this batch and x_n is the predict of this batch.

Through this loss function, we can use radiologists' domain knowledge to guide the TAM focusing on important time slots. Besides, as with TAM, DKG-TAM is also added after the third 3D convolutional layer of the backbone.

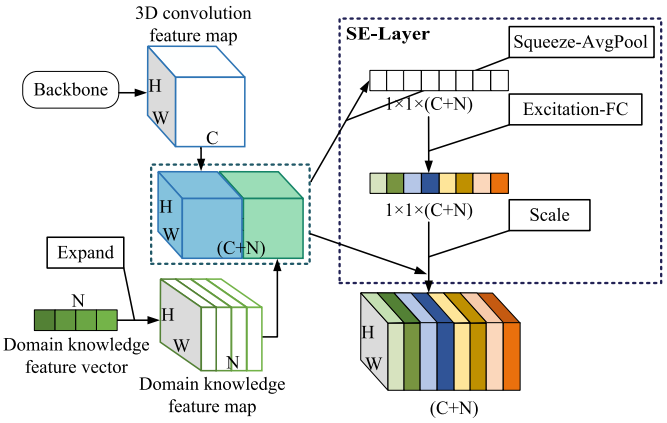


Fig. 6. The structure of DKG-CAM. The DKG-CAM concatenates the features extracted by C3D and the features extracted based on radiologists' domain knowledge in the channel dimension. The SE-Layer in DKG-CAM is used to reallocate the weights of each feature.

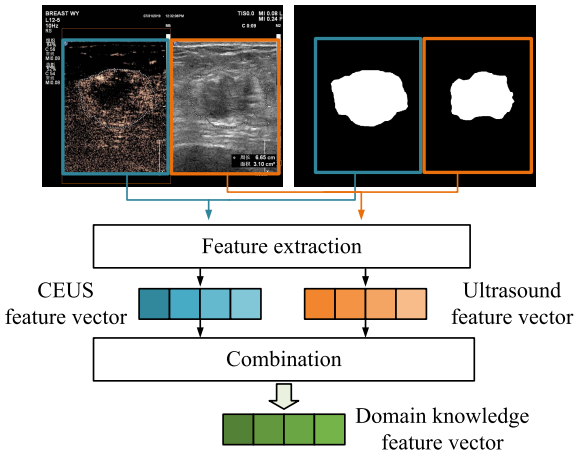


Fig. 7. The extraction of domain knowledge feature vector. First, the CEUS feature vector and US feature vector are calculated by the corresponding mask and original image. Then the domain knowledge vector is calculated from the ratio of CEUS feature vector and US feature vector.

D. Domain-Knowledge-Guided Channel Attention Module

Besides the temporal diagnostic pattern mentioned above, there is another pattern in the diagnosis. Radiologists usually focus on the difference in tumor diameter between the CEUS window and the US window, especially at the moment when the tumor boundary is most clearly shown in a CEUS window. Therefore, we build a domain-knowledge-guided channel attention module (DKG-CAM). The specific structure of the DKG-CAM is shown as Fig. 6.

Firstly, we perform feature extraction separately for the two windows in the frame chosen by radiologists. The frame is chosen from the whole CEUS video with the clearest boundary of the tumor in the CEUS window. Through correlation analysis between features and classification results on the training set, we select 32 features in total, including shape features, first-order features, gray level size zone matrix features, and others. The details of the extracted features are shown in Table II. First-order features describe the voxel intensity distribution of the input regions. The shape features include descriptors of the two-dimensional size and shape of the region of interest. Gray level size zone matrix (GLSZM) calculates gray level zones in the region of interest. Gray level dependence matrix (GLDM) quantifies gray level dependencies in the region of interest. Gray level run length matrix (GLRLM) describes gray level runs, which are defined as the length in the number of pixels, of consecutive pixels that have the same gray level value. Neighbouring gray tone difference matrix (NGTDM) quantifies the difference between a gray value and the average gray value of its neighbours within distance δ.

We then use the feature vectors extracted in CEUS and US windows to calculate the ratio and obtain the domain knowledge feature vector. This process can be seen in Fig. 7.

We expand the domain knowledge feature vector to a domain knowledge feature map of shape $32\times3\times4$, and concatenate the domain knowledge feature map with the feature map output from the fourth 3D convolutional layer. Then we use the SE-Layer to allocate attention to different channels, which can selectively emphasize informative features and suppress less useful ones.

IV. Dataset and Experiments

A. Breast-CEUS Dataset

Although CEUS has been widely used in breast cancer diagnosis, there are no public breast CEUS datasets available in CAD. To advance research in this field, we collect breast CEUS video data from the Cancer Hospital and Institute, Peking Union Medical College and Chinese Academy of Medical Science (CICAMS), and construct a Breast-CEUS dataset. The study population consisted of 221 breast lesions in 217 patients who were examined at CICAMS between March 2019 and November 2020 and had a preliminary clinical diagnosis of breast cancer. Four of the patients provided eight cases because they had tumors in both the right and left breasts. All patients were treated surgically at CICAMS and completed postoperative pathological tests. Patients were given US and CEUS examinations before surgery. Consent was obtained from the participants, and the study was approved by the ethics committee of the participating hospital.

Inclusion Criteria:

- Patients with a preliminary clinical diagnosis of mammary tumor who underwent surgery in CICAMS and obtained complete pathological results;
- Patients who have voluntarily cooperated in the US and CEUS examinations;

$$L_{tmp}(trend_{radiologist}, trend_{model}) = -\log \frac{\sum_{i=1}^{n} trend_{radiologist,i} \times trend_{model,i}}{\sqrt{\sum_{i=1}^{n} trend_{radiologist,i}^2} \times \sqrt{\sum_{i=1}^{n} trend_{model,i}^2}} \quad (2)$$

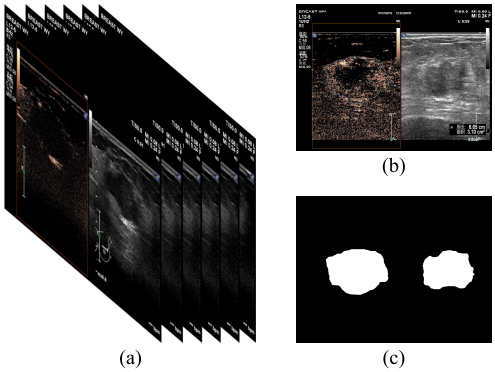


Fig. 8. One case of the Breast CEUS Dataset. (a) All frames of a CEUS video. (b) The dual-amplitude CEUS frame chosen by the radiologists from (a), which can show a clear boundary of the tumor in the CEUS window. (c) Mask of tumors in the CEUS window and US window in (b).

- Patients who have not receive a gross needle biopsy of the breast or other treatment before the US examination.

Exclusion criteria:

- Patients who have local skin ulceration and cannot complete the US examination;
- Patients who are allergic to the US contrast agent and have contraindications to the US such as severe cardiopulmonary function, liver and kidney dysfunction;
- Patients who refuse to sign the informed consent form for US examinations;
- Patients who have received a gross needle biopsy of the breast or other treatment for breast cancer.

The detailed process of US examination:

1) Selecting high-frequency line array probes (frequency 5-12MHz) and breast imaging conditions in the Philips EPIQ5 ultrasound instrument (Philips Bothell, WA).
2) Focusing on the center of the tumor, and then sweep the target tumor in multiple sections and angles to capture images of the largest section of the tumor.

The detailed process of CEUS examination:

1) Injecting 5 ml 0.9% NaCl solution into SonoVue and shake to obtain a microbubble suspension.
2) Using Philips EPIQ5 diagnostic ultrasound instrument, selecting a high-frequency line array probe (5-12 Mhz) and switching the instrument to real-time dual-amplitude contrast-enhanced ultrasound imaging mode.
3) Injecting the prepared contrast agent into the peripheral venous.
4) Starting the image storage function simultaneously, and record the microcirculatory perfusion of the tumor for 3 minutes.

Each case in the dataset contains a CEUS video, a dual-amplitude CEUS image, and a mask corresponding to the CEUS image, as shown in Fig. 8. The CEUS video resolution is 1024×768, and each video is 3 minutes with nearly 5 FPS. Fig. 8 (a) is a frame from the CEUS video. The CEUS video contains two windows, the left window is the CEUS part, and the right window is the US part. The tumor positions are the same in both windows, and the two windows are obtained by sampling at the same time. The dual-amplitude CEUS image is the frame chosen by the radiologist from the CEUS video with the most clear boundary of the tumor in the CEUS window. The mask corresponding to the CEUS image is obtained through the radiologist outlining the boundary of the tumor in the CEUS and US windows, respectively.

Complete surgical pathology results are available for all cases in the dataset. Efforts are being made to make the Breast-CEUS dataset open-access for wider use internationally by researchers.

B. Exprimental Settings

1) Evaluation: We evaluate our method on the Breast-CEUS dataset. Since the imbalance of categories in our dataset (less benign video data), we implement data augmentation on part of the benign data. The procedure for data augmentation on CEUS videos of benign cases is to perform a horizontal flip of the US and CEUS portions of each frame of the video, respectively. After augmentation, we acquire the Breast-CEUS dataset consisting of 270 CEUS video cases (i.e., 150 malignant and 120 benign).

Considering the limited size of the Breast CEUS Dataset, we apply ten-fold cross-validation on the augmented dataset. The samples contained in each fold are non-overlapping and mutually exclusive. We use Sensitivity (Se), Specificity (Sp), Accuracy (Acc), F_β-score (F_β) as the primary evaluation metrics [46]–[48].

2) Settings: We implement our model in Pytorch. Besides, we use NVIDIA-APEX to accelerate the model training process. In our experiments, the batch size is set to 64 for 60 epochs, with the learning rate of 1e-4. To mitigate overfitting, we also use dropout layers in our model, and the dropout rate is set to 0.1. The whole model is randomly initialized by Pytorch with no pretraining on any other dataset.

During training, we set the input of the model to $100 \times 1 \times 240 \times 320$, which means that we sample 100 frames at equal time intervals from the whole CEUS video, convert each frame to a grayscale image, and resize each to 240×320. Besides, each frame of the video is standardized. All experiments are conducted on a server containing two GPUs of NVIDIA Tesla V100 32GB.

C. Exprimental Results

To demonstrate the effectiveness of our proposed method, we conduct a series of comparison and ablation experiments. Besides, we analyze the attention weights obtained from DKG-TAM and DKG-CAM. The details of the experimental results are as follows.

1) Overall Comparison and Ablation Experiment Results: We design different comparison experiments to evaluate the validity of the proposed method. To demonstrate the role of CEUS videos and domain knowledge, we design three sets of experiments: the first set of experiments use only ultrasound images, the second set of experiments use only CEUS videos, and the third set of experiments use both ultrasound images and CEUS videos in combination with domain knowledge. The results are summarized in Table III. In particular, underlined

TABLE III
THE PERFORMANCE COMPARISON OF SEVEN METHODS ON THE US AND CEUS DATASET. THE ABLATION RESULTS OF DKG-TAM AND DKG-CAM ON THREE METHODS ARE SHOWN AT THE BOTTOM OF THE TABLE

	DKG TAM	DKG CAM	Data Modility	$Se(95\%CI)$	$Sp(95\%CI)$	$Acc(95\%CI)$	$F_{0.5}$	F_1	F_2	P
ResNet-50 [49]			US	76.7% (68.9,83.0)	64.2% (54.9,72.6)	71.1% (65.4,76.2)	73.5%	74.7%	75.9%	-
ResNeXt-50 [50]			US	82.0% (74.7,87.6)	68.3% (59.1,76.4)	75.9% (70.5,80.7)	77.5%	79.1%	80.8%	-
Inception v3 [51]			US	83.3% (76.2,88.7)	70.0% (60.9,77.9)	77.4% (72.0,82.0)	78.7%	80.4%	82.1%	-
C3D [52]			CEUS	85.3% (78.4,90.4)	69.2% (60.00,77.1)	78.2% (72.8,82.7)	79.0%	81.2%	93.7%	-
R3D [53]			CEUS	84.7% (77.8,89.9)	70.8% (61.7,78.6)	78.5% (73.2,83.0)	79.6%	81.4%	83.3%	-
TRN [54]			CEUS	88.7% (82.2,93.1)	68.33% (59.1,76.4)	79.6% (74.4,84.0)	79.7%	82.9%	86.3%	-
Ours			CEUS	93.7% (92.3,95.2)	66.5% (64.9,68.1)	81.6% (81.1,82.1)	80.5%	85.0%	90.0%	<0.001
C3D	√		CEUS	86.7% (79.9,91.5)	70.8% (61.7,78.6)	79.6% (74.4,84.0)	80.3%	82.5%	85.0%	-
		√	US+CEUS	86.0% (79.2,90.9)	72.5% (63.5,80.1)	80.0% (74.8,84.4)	80.8%	82.7%	84.7%	-
	√	√	US+CEUS	86.7% (79.9,91.5)	74.2% (65.2,81.5)	81.1% (76.0,85.4)	81.9%	83.6%	85.4%	-
R3D	√		CEUS	83.3% (76.2,88.7)	**77.5%** (68.8,84.4)	80.7% (75.6,85.0)	82.5%	82.8%	83.1%	-
		√	US+CEUS	84.0% (76.9,89.3)	**75.0%** (66.1,82.3)	80.0% (76.0,86.3)	81.4%	82.4%	83.3%	-
	√	√	US+CEUS	92.7% (86.9,96.1)	72.5% (63.5,80.1)	83.7% (78.8,87.7)	**82.9%**	86.3%	90.0%	-
Ours	√		CEUS	**96.3%** (95.4,97.1)	68.1% (67.2,69.0)	83.7% (83.3,84.2)	82.0%	86.8%	**92.2%**	<0.001
		√	US+CEUS	94.8% (94.3,95.3)	70.8% (70.1,71.5)	**84.2%** (83.8,84.5)	82.8%	**86.9%**	91.5%	<0.001
	√	√	US+CEUS	**97.2%** (96.7,97.7)	72.5% (71.7,73.3)	**86.3%** (85.9,86.6)	**84.3%**	**88.7%**	**93.6%**	<0.001

and bolded results are the highest of their kind, and bolded only results are the second-highest of their kind.

In the first set of experiments, we use some currently popular image classification models to classify the US images in our dataset, such as ResNet [49], ResNeXt [50] and Inception [51]. In the second set of experiments, we choose classical methods including C3D [52] and R3D [53] and the newly proposed method TRN [54] for comparison. We choose C3D as our baseline because our proposed vanilla network is based on C3D. From the comparison results, we can see that the CEUS videos have some advantages over the US images in breast tumor classification. In our dataset, the best accuracy with ultrasound images is 77.4% obtained by Inception v3, while the best accuracy with CEUS videos is 81.6% obtained by our method. Our vanilla network outperforms the baseline in all metrics except specificity and F_2 score. The third set of experiments are ablation experiments of DKG-TAM and DKG-CAM. The results show that our model can achieve the highest accuracy of 86.3% when using DKG-TAM and DKG-CAM for multimodality information fusion. Also, the results of the ablation experiments of these two modules on different models (such as C3D and R3D) demonstrate the effectiveness and robustness of these two modules. The overall results demonstrate that the domain knowledge module we designed is useful for breast tumor classification tasks and can adequately incorporate multimodality information. In ablation experiments for both DKG-TAM and DKG-CAM on our method, we perform ten repeated experiments, retaining the mean values of each metric, and we use the student's t test for statistical analysis. Since their p values are less than 0.001, these experimental results all reach statistical significance.

To better understand the effects of DKG-TAM and DKG-CAM, we investigate the influence of each module on the performance of our model. As shown in Table III, by adding DKG-TAM and DKG-CAM to our method respectively, the accuracy can get the boost of 2.1% and 2.6%. In addition,

TABLE IV
EFFECTIVENESS OF THE PEAK POSITION OF THE BRIGHTNESS CHANGE CURVE ON OUR METHOD

Peak Position	Se	Sp	Acc	$F_{0.5}$	F_1	F_2
The second time slot	**97.2%**	**72.5%**	**86.3%**	**84.3%**	**88.7%**	**93.6%**
The third time slot	94.0%	**71.7%**	84.1%	82.9%	86.8%	91.0%
The fourth time slot	**97.3%**	70.0%	85.2%	83.1%	88.0%	93.4%
Mixed	**97.3%**	70.8%	**85.6%**	**83.5%**	**88.2%**	**93.5%**
P			<0.001			

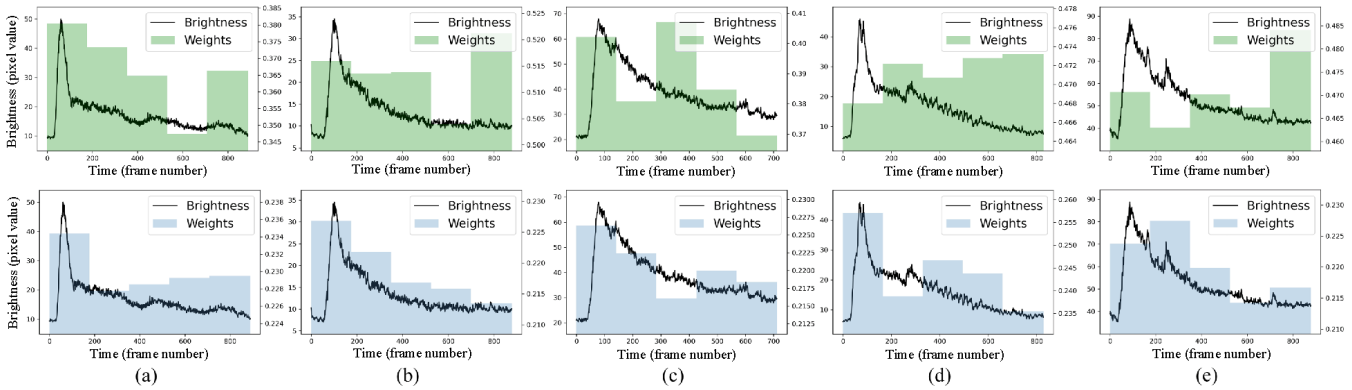


Fig. 9. (a)-(e) represent five different cases that are misclassified by TAM but correctly classified by DKG-TAM. The comparison of the top and bottom rows demonstrate that DKG-TAM has learned temporal attention that better matches the brightness change curve. This higher quality of temporal attention improves the accuracy of the classification. The first row shows the temporal attention of TAM in these cases. The second row shows the temporal attention of DKG-TAM in these cases. The histogram shows the temporal attention generated by TAM, and the line graph shows the brightness change curve of the input video.

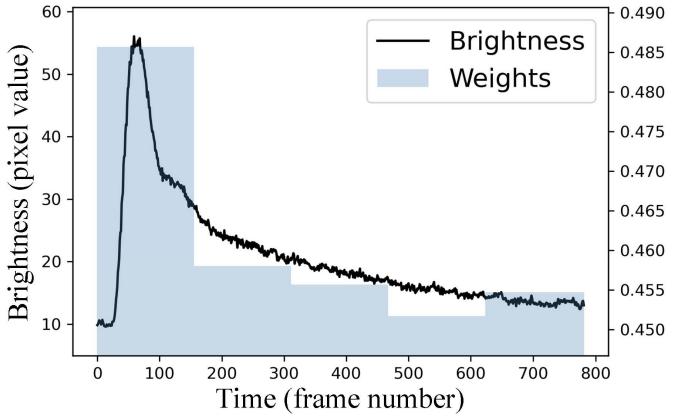


Fig. 10. Temporal attention and brightness changes for one of the correctly classified cases. This figure shows that the temporal attention of TAM and the brightness change curve match well on correctly classified cases. The histogram shows the temporal attention generated by TAM, and the line shows the brightness change curve of the input video.

adding both DKG-TAM and DKG-CAM simultaneously can further improve the performance and get the best accuracy as 86.3%.

According to the results, by adding DKG-TAM, we can achieve a 2.6% gain on sensitivity while the specificity improves 1.6% compared with our vanilla network. The addition of DKG-CAM can achieve a specificity of 70.8%, delivering a gain of 4.3% compared with our vanilla network, and a 1.1% improvement of sensitivity. The increase in specificity suggests that DKG-CAM reduces the number of false positives in the model and is better to identify tumors.

Finally, we implement our model with both DKG-TAM and DKG-CAM into our Breast-CEUS dataset. It achieves the highest sensitivity of 97.2% and the highest accuracy of 86.3%. With respect to other metrics, our model also achieves optimal or suboptimal results. From the results, the effects of the two modules seem to be orthogonal. The reason might be that the DKG-TAM and DKG-CAM help the model improve performance from completely different aspects (temporal and channel dimension).

In addition, we apply DKG-CAM and DKG-TAM to C3D and R3D, respectively, the accuracy gain 2.9% and 5.2% improvements, further proving the effectiveness of our domain knowledge modules.

Since the peaks of the brightness curves of CEUS videos mostly fall into the second time slot, we design the following experiments to verify whether our model is sensitive to the position of the signals' peak. First, we truncate the original CEUS videos so that the maximum peaks of the signal fall into the third or the fourth time slot. We use the proposed method on the truncated videos for training and testing, and the results are shown in Table IV. The results show that the accuracy floats less than 2% no matter which time slot the peaks fall into. In the statistical analysis of the experimental results, the p value is less than 0.001 which indicates that our method is statistically insensitive to the peaks' position. This further demonstrates that our method learns the radiologists' domain knowledge of temporal.

TABLE V
EFFECTIVENESS OF TAM AND TEMPORAL ATTENTION FROM DOMAIN KNOWLEDGE ON OUR METHOD

TAM	Domain Knowledge		Se	Sp	Acc	$F_{0.5}$	F_1	F_2
	mean	range						
			93.7%	**66.5%**	81.6%	80.5%	85.0%	90.0%
√			96.7%	65.8%	**83.0%**	**81.1%**	**86.3%**	92.2%
	√		96.7%	64.2%	82.2%	80.4%	85.8%	92.0%
		√	**98.0%**	59.2%	80.7%	78.7%	85.0%	**92.3%**
√	√		96.3%	**68.1%**	**83.7%**	**82.0%**	**86.8%**	92.2%
√		√	**98.7%**	59.2%	81.1%	78.9%	85.3%	**92.9%**

TABLE VI
EFFECTIVENESS OF THE TIME SLOT WIDTH USED BY DKG-TAM ON OUR METHOD

Position	Time slot width (frame number)	Se	Sp	Acc	$F_{0.5}$	F_1	F_2
After 3dConv_1	2	96.0%	62.5%	81.1%	79.5%	85.0%	91.3%
After 3dConv_2	4	**96.7%**	**64.2%**	**82.2%**	**80.4%**	**85.8%**	**92.0%**
After 3dConv_3	20	**96.3%**	**68.1%**	**83.7%**	**82.0%**	**86.8%**	**92.2%**

2) Ablation Experiment Results on TAM: In order to verify the contribution of TAM and the domain knowledge in terms of performance improvement, we also perform more detailed ablation experiments on our method, as shown in Table V. In the ablation experiments, TAM uses a series of convolutional layers to compute temporal attention to weigh the feature map. Domain knowledge denotes that weights the feature map using the temporal attention from radiologists' domain knowledge. Mean and range in Table V denote the calculation of mean and range respectively for the brightness change in CEUS video. When adding TAM to the model, the sensitivity reaches 96.7%, but the specificity is only 65.8%. Overall, compared with our vanilla network, the addition of TAM improves the sensitivity of 3.0% and accuracy of 1.4%.

Besides, we analyze the weights of TAM. Fig. 9 shows the temporal attention obtained from TAM for some cases. We notice that the temporal attention of the cases classified correctly among them largely matches the trend of the brightness change curve, as shown in Fig. 10. Since the brightness change responds to radiologists' attention, this result demonstrates our previous assumption that the accuracy can be improved if the model can focus on the time slots that radiologists pay attention to.

Firstly, we directly apply radiologists' temporal attention to the feature map. In the acquisition of the radiologists' temporal attention from domain knowledge, we use the methods of calculating the mean and calculating the range, which obtains sensitivities of 96.7% and 98.0% respectively. These two methods calculate the mean or the range for all frames in a time slot to represent the temporal attention of that time slot. The significant improvement of calculating the mean is because it takes full advantage of all the frames in a time slot.

Then we integrate DKG-TAM into the model, which contains both TAM and domain knowledge. This model calculating the mean of brightness to constrain the temporal attention in TAM achieves the sensitivity of 96.3% and the highest accuracy of 83.7%. The other metrics of this method mostly achieve optimal or suboptimal results, respectively. In contrast, the model calculating the range of brightness to constrain the temporal attention in TAM achieves specificity of 59.2% and accuracy of 81.1%. This result proves the superiority of calculating the mean over calculating the range of brightness. Compared with adding TAM or using radiologists' attention directly, DKG-TAM uses radiologists' attention to constrain the temporal attention obtained by TAM. This mechanism allows the module to learn the trend of radiologists' attention allocation while also automatically adjusting for the difference in attention allocation over different temporal periods, resulting in the optimal results for DKG-TAM.

In addition, we test the performance of DKG-TAM with different widths of time slots, and the results are shown in Table VI. The results show that the accuracy decreases as the time slots used by DKG-TAM become slimmer. We consider the reason is that slimmer time slots lead to a lack of high-level features in the feature maps used by DKG-TAM. Since the slimmer time slots exist in the feature maps from the shallow layers of the neural network, and the feature maps extracted from the shallow layers lack high-level features.

3) Ablation Experiment Results on CAM: To demonstrate that the addition of domain knowledge features and the contribution of CAM to model performance improvement, we do ablation experiments on our method, and the results are shown in Table VII. In the ablation experiments, CAM indicates that we do not add domain knowledge features and apply the SE-Layer directly to the feature map. Domain knowledge indicates that we concatenate the extracted domain knowledge features directly to the feature map without reweighting. When CAM is added to the model, sensitivity increase by 5.0% compared with our vanilla network, and achieves the highest.

TABLE VII
EFFECTIVENESS OF CAM AND DOMAIN KNOWLEDGE FEATURES ON OUR METHOD

CAM	Domain Knowledge	Se	Sp	Acc	$F_{0.5}$	F_1	F_2
		93.7%	66.5%	81.6%	80.5%	85.0%	90.0%
√		**98.7%**	64.2%	83.3%	81.0%	**86.8%**	**93.6%**
	√	94.7%	**70.0%**	**83.7%**	**82.4%**	86.6%	91.3%
√	√	**94.8%**	**70.8%**	**84.2%**	**82.8%**	**86.9%**	**91.5%**

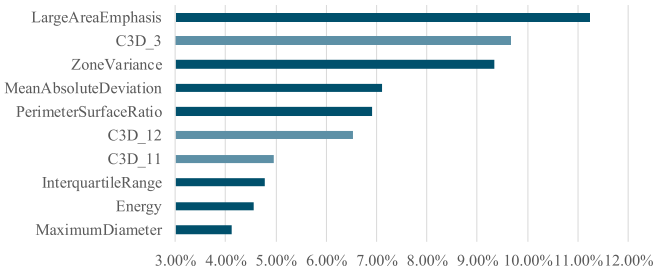


Fig. 11. The average weight of CAM for all cases. The prefix C3D means the features are extracted by C3D backone, and other features are extracted based on the radiologists' domain knowledge. A higher value of weight indicates the feature plays a more important role in classification.

When domain knowledge features are added to the model alone, sensitivity increased by 1.0% and specificity increased by 3.5% compared with our vanilla network. When both CAM and domain knowledge features are added to the model, sensitivity achieves the second highest score of 94.8% and specificity achieves the highest score of 70.8%. Both accuracy and F_1-score also achieve the highest when adding DKG-CAM. The results show that the CAM can pay more attention to informative features, and the addition of domain knowledge can increase the number of valid features that used by the model. DKG-CAM pays more attention to features that are used to distinguish true negative cases from false positive cases.

4) Analysis of the Weights of the Attention Module: In addition to analyzing these evaluation metrics, we also analyze the attention weights of DKG-TAM and DKG-CAM.

As can be seen in Fig. 9, DKG-TAM is able to classify misclassified cases in TAM correctly by constraining its temporal attention toward the radiologists' temporal attention distribution. This result is further evidence of the contribution of radiologists' domain knowledge incorporation to model performance.

Then we analyze the weights of the DKG-CAM. The weights generated by the DKG-CAM is a vector of 56 dimensions. Since the DKG-CAM is self-learning, this weight vector represents the importance of these 56 features (24 features extracted by the C3D backbone and another 32 features extracted manually based on the radiologists' domain knowledge) in the classification tasks. We select the top ten features in the weight vector for display, as shown in Fig. 11. From this figure, seven of the top ten features that are important for classification are domain knowledge features we added. The phenomenon that the perimeter surface ratio and the maximum diameter of tumors ranking ahead is also consistent with the radiologists' domain knowledge. It also indicates that there is no great overlap between our extracted domain knowledge features and those extracted by the C3D backbone and that our domain knowledge features can effectively improve the model's ability to classify tumors and play an important role compared with those extracted by C3D.

By analyzing the weighting of the two modules, we believe that the addition of these two modules allowed the model to learn radiologists' diagnostic pattern of CEUS videos and the features radiologists interested in the diagnostic process.

D. Discussion

1) Performance: First, we discuss why the accuracy of our method is better than that of C3D and R3D. Due to the small size of the Breast-CEUS dataset, we argue that the overfitting of C3D and R3D on our dataset is unavoidable. Although R3D has residual modules, the residual module is designed to solve the gradient disappearance problem during training, so it does not mitigate the overfitting well. Our model based on C3D can mitigate the overfitting by the reduced network layers. Therefore, our method exceeds the accuracy of C3D and R3D by 3.4% and 3.1%, without the addition of DKG-CAM and DKG-TAM. When both adding DKG-CAM and DKG-TAM, our method has 5.2% or 2.6% advantages in accuracy compared to C3D and R3D.

In the comparison experiments with C3D and R3D, we notice that when adding DKG-CAM and DKG-TAM to C3D and R3D respectively, the performance of both C3D and R3D could be improved, especially for R3D, which achieved the highest 5.2% improvement in accuracy. We believe that this is due to the residual structure in R3D, which results in R3D retaining enough detail in the back layer of the network compared to C3D. The DKG-TAM and DKG-CAM capture these details, resulting in a better improvement for the model.

In addition, we would like to discuss the observation that sensitivity is generally higher than specificity in our experimental results. We think there are two reasons for this observation. On one hand, we believe it is caused by the imbalance in the Breast-CEUS dataset. On the other hand, during breast cancer diagnosis, radiologists focus more on malignant cases. Therefore, the domain knowledge we integrated is mostly related to the diagnosis of malignancy.

2) Scope of Application: In this paper, we build several modules, such as TAM, DKG-TAM, and DKG-CAM. The TAM can be inserted anywhere in any C3D model, and this module can provide an improvement in the performance of the model with a low complexity increase. DKG-TAM is a supervised module that requires additional annotation. The temporal attention annotation of the data is difficult to obtain, which limits the scope of DKG-TAM. DKG-CAM introduces extracted domain knowledge features and automatically allocates the weights of the fused feature map in the channel dimension, providing a new way of adding handmade features to deep models. The module can be extended easily to other models that incorporate handmade features. In addition, since our dataset is collected from one scanner, there may be performance degradation when our model is applied to data collected from different scanners.

3) Future Work: Since the size of Breast-CEUS dataset is not large, we are collecting more CEUS videos currently. In addition, another interesting research point is to explore more types of data to improve the performance of neural networks (e.g. tissue biomarkers from elastography). Considering that the data in our dataset are collecting from one scanner, we will collaborate with more hospitals to collect videos from different scanners to complement our dataset. In addition, for the actual diagnostic process, we will make the model light-weighted to be able to implement in real condition. Furthermore, we will develop a CAD system based on our model, which can give classification results in a real-time manner during the CEUS examination of radiologists.

V. Conclusion

In this paper, we incorporate radiologists' domain knowledge and present a CEUS video classification model for breast cancer. The model consists of 3D convolution with a temporal attention module incorporating temporal domain knowledge and a channel attention module that incorporates feature-based domain knowledge. With two domain knowledge modules, our model is capable of focusing on critical time slots of CEUS videos and extract features more efficiently, which helps to improve the classification performance of the model. The final model achieves a sensitivity of 97.2% and an accuracy of 86.3% on a dataset containing 221 cases. By integrating the domain knowledge of radiologists, our approach improves the classification ability of deep learning models, ultimately achieving the best results on our dataset.

Currently, CEUS is becoming increasingly important in the diagnosis of breast cancer by radiologists. In this paper, we explore the feasibility of applying neural networks to CEUS data for breast cancer diagnosis and design two attention modules to effectively integrate radiologists' domain knowledge into neural networks. We also use the weights of the attention module to identify the improvement of domain knowledge on breast cancer diagnosis.

Acknowledgment

This research is based on the CEUS data collected by Cancer Hospital Chinese Academy of Medical Sciences, and thanks to the doctors of this hospital for their help throughout the research process. The authors would also like to thank Chen Zhengsu, Xie Xiaozheng, Meng Hui for their valuable suggestions.

References

[1] A. Das, U. R. Acharya, S. S. Panda, and S. Sabut, "Deep learning based liver cancer detection using watershed transform and Gaussian mixture model techniques," *Cognit. Syst. Res.*, vol. 54, pp. 165–175, May 2019.

[2] J. R. Burt *et al.*, "Deep learning beyond cats and dogs: Recent advances in diagnosing breast cancer with deep neural networks," *Brit. J. Radiol.*, vol. 91, no. 1089, Apr. 2018, Art. no. 20170545.

[3] X. Liu *et al.*, "Localization and diagnosis framework for pediatric cataracts based on slit-lamp images using deep features of a convolutional neural network," *PLoS ONE*, vol. 12, no. 3, Mar. 2017, Art. no. e0168606.

[4] R. Rasti, M. Teshnehlab, and S. L. Phung, "Breast cancer diagnosis in DCE-MRI using mixture ensemble of convolutional neural networks," *Pattern Recognit.*, vol. 72, pp. 381–390, Dec. 2017.

[5] B. Xu *et al.*, "Attention by selection: A deep selective attention approach to breast cancer classification," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 1930–1941, Jun. 2020.

[6] L. Luo *et al.*, "Deep angular embedding and feature correlation attention for breast MRI cancer analysis," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Springer, 2019, pp. 504–512.

[7] L. Liberman, T. L. Feng, D. D. Dershaw, E. A. Morris, and A. F. Abramson, "US-guided core breast biopsy: Use and cost-effectiveness," *Radiology*, vol. 208, no. 3, pp. 717–723, Sep. 1998.

[8] W. A. Berg *et al.*, "Combined screening with ultrasound and mammography vs mammography alone in women at elevated risk of breast cancer," *J. Amer. Med. Assoc.*, vol. 299, no. 18, pp. 2151–2163, 2008.

[9] P. Kijanka and M. W. Urban, "Local phase velocity based imaging: A new technique used for ultrasound shear wave elastography," *IEEE Trans. Med. Imag.*, vol. 38, no. 4, pp. 894–908, Apr. 2019.

[10] S. Goswami, R. Ahmed, S. Khan, M. M. Doyley, and S. A. McAleavey, "Shear induced non-linear elasticity imaging: Elastography for compound deformations," *IEEE Trans. Med. Imag.*, vol. 39, no. 11, pp. 3559–3570, Nov. 2020.

[11] P. Kapetas *et al.*, "Quantitative multiparametric breast ultrasound: Application of contrast-enhanced ultrasound and elastography leads to an improved differentiation of benign and malignant lesions," *Investigative Radiol.*, vol. 54, no. 5, pp. 257–264, May 2019.

[12] C. F. Wan, J. Du, H. Fang, F. H. Li, J. S. Zhu, and Q. Liu, "Enhancement patterns and parameters of breast cancers at contrast-enhanced US: Correlation with prognostic factors," *Radiology*, vol. 262, no. 2, pp. 450–459, Feb. 2012.

[13] C.-F. Wan, X.-S. Liu, L. Wang, J. Zhang, J.-S. Lu, and F.-H. Li, "Quantitative contrast-enhanced ultrasound evaluation of pathological complete response in patients with locally advanced breast cancer receiving neoadjuvant chemotherapy," *Eur. J. Radiol.*, vol. 103, pp. 118–123, Jun. 2018.

[14] K. Cox *et al.*, "Validation of a technique using microbubbles and contrast enhanced ultrasound (CEUS) to biopsy sentinel lymph nodes (SLN) in pre-operative breast cancer patients with a normal grey-scale axillary ultrasound," *Eur. J. Surgical Oncol.*, vol. 39, no. 7, pp. 760–765, Jul. 2013.

[15] D. Tran, H. Wang, M. Feiszli, and L. Torresani, "Video classification with channel-separated convolutional networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5552–5561.

[16] L. Wang, P. Koniusz, and D. Huynh, "Hallucinating IDT descriptors and I3D optical flow features for action recognition with CNNs," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8698–8708.

[17] Y. Du, C. Yuan, B. Li, L. Zhao, Y. Li, and W. Hu, "Interaction-aware spatio-temporal pyramid attention networks for action classification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 373–389.

[18] S. Li, S. Bak, P. Carr, and X. Wang, "Diversity regularized spatiotemporal attention for video-based person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 369–378.

[19] J. Carreira, E. Noland, A. Banki-Horvath, C. Hillier, and A. Zisserman, "A short note about kinetics-600," 2018, *arXiv:1808.01340*. [Online]. Available: http://arxiv.org/abs/1808.01340

[20] K. Soomro, A. Roshan Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," 2012, *arXiv:1212.0402*. [Online]. Available: http://arxiv.org/abs/1212.0402

[21] Z. Zhu, W. Wu, W. Zou, and J. Yan, "End-to-end flow correlation tracking with spatial-temporal attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Feb. 2018, pp. 548–557.

[22] Y.-X. Zhao, S. Liu, Y.-B. Hu, Y.-Y. Ge, and D.-M. Lv, "Diagnostic and prognostic values of contrast-enhanced ultrasound in breast cancer: A retrospective study," *OncoTargets therapy*, vol. 10, p. 1123, Jul. 2017.

[23] Y. Zhang, B. Zhang, X. Fan, and D. Mao, "Clinical value and application of contrast-enhanced ultrasound in the differential diagnosis of malignant and benign breast lesions," *Experim. Therapeutic Med.*, vol. 4, pp. 2063–2069, Jun. 2020.

[24] R. Guo, G. Lu, B. Qin, and B. Fei, "Ultrasound imaging technologies for breast cancer detection and management: A review," *Ultrasound Med. Biol.*, vol. 44, no. 1, pp. 37–70, Jan. 2018.

[25] A. Sridharan, J. R. Eisenbrey, J. K. Dave, and F. Forsberg, "Quantitative nonlinear contrast-enhanced ultrasound of the breast," *Amer. J. Roentgenol.*, vol. 207, no. 2, pp. 274–281, Aug. 2016.

[26] C. Zhang, Z. Yang, X. He, and L. Deng, "Multimodal intelligence: Representation learning, information fusion, and applications," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 3, pp. 478–493, Mar. 2020.

[27] B. Nojavanasghari, D. Gopinath, J. Koushik, T. Baltrušaitis, and L.-P. Morency, "Deep multimodal fusion for persuasiveness prediction," in *Proc. 18th ACM Int. Conf. Multimodal Interact.*, 2016, pp. 284–288.

[28] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear CNN models for fine-grained visual recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1449–1457.

[29] Q. Zhang *et al.*, "Deep learning based classification of breast tumors with shear-wave elastography," *Ultrasonics*, vol. 72, pp. 150–157, Dec. 2016.

[30] S. Han *et al.*, "A deep learning framework for supporting the classification of breast lesions in ultrasound images," *Phys. Med. Biol.*, vol. 62, no. 19, p. 7714, 2017.

[31] S. Duraisamy and S. Emperumal, "Computer-aided mammogram diagnosis system using deep learning convolutional fully complex-valued relaxation neural network classifier," *IET Comput. Vis.*, vol. 11, no. 8, pp. 656–662, Dec. 2017.

[32] G. Carneiro, J. Nascimento, and A. P. Bradley, "Unregistered multiview mammogram analysis with pre-trained deep learning models," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Springer, 2015, pp. 652–660.

[33] R. K. Samala, H.-P. Chan, L. M. Hadjiiski, M. A. Helvie, K. H. Cha, and C. D. Richter, "Multi-task transfer learning deep convolutional neural network: Application to computer-aided diagnosis of breast cancer on mammograms," *Phys. Med. Biol.*, vol. 62, no. 23, p. 8894, 2017.

[34] H. Lee, J. Park, and J. Y. Hwang, "Channel attention module with multiscale grid average pooling for breast cancer segmentation in an ultrasound image," *IEEE Trans. Ultrason., Ferroelectr., Freq. Control*, vol. 67, no. 7, pp. 1344–1353, Jul. 2020.

[35] M. Byra, H. Piotrzkowska-Wroblewska, K. Dobruch-Sobczak, and A. Nowicki, "Combining nakagami imaging and convolutional neural network for breast lesion classification," in *Proc. IEEE Int. Ultrason. Symp. (IUS)*, Sep. 2017, pp. 1–4.

[36] Z. Gandomkar, P. C. Brennan, and C. Mello-Thoms, "MuDeRN: Multi-category classification of breast histopathological image using deep residual networks," *Artif. Intell. Med.*, vol. 88, pp. 14–24, Jun. 2018.

[37] Y. Feng, L. Zhang, and Z. Yi, "Breast cancer cell nuclei classification in histopathology images using deep neural networks," *Int. J. Comput. Assist. Radiol. Surgery*, vol. 13, no. 2, pp. 179–191, Feb. 2018.

[38] J.-Z. Cheng *et al.*, "Computer-aided diagnosis with deep learning architecture: Applications to breast lesions in US images and pulmonary nodules in CT scans," *Sci. Rep.*, vol. 6, no. 1, Apr. 2016, Art. no. 24454.

[39] O. Hadad, R. Bakalo, R. Ben-Ari, S. Hashoul, and G. Amit, "Classification of breast lesions using cross-modal deep learning," in *Proc. IEEE 14th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2017, pp. 109–112.

[40] X. Zhang *et al.*, "Whole mammogram image classification with convolutional neural networks," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Oct. 2017, pp. 700–704.

[41] D.-A. Clevert, A. Horng, D.-A. Clevert, E. M. Jung, W. H. Sommer, and M. Reiser, "Contrast-enhanced ultrasound versus conventional ultrasound and MS-CT in the diagnosis of abdominal aortic dissection," *Clin. Hemorheol. Microcirculat.*, vol. 43, nos. 1–2, pp. 129–139, 2009.

[42] A. Nielsen Moody *et al.*, "Preoperative sentinel lymph node identification, biopsy and localisation using contrast enhanced ultrasound (CEUS) in patients with breast cancer: A systematic review and meta-analysis," *Clin. Radiol.*, vol. 72, no. 11, pp. 959–971, Nov. 2017.

[43] Z. Yang, X. Gong, Y. Guo, and W. Liu, "A temporal sequence dual-branch network for classifying hybrid ultrasound data of breast cancer," *IEEE Access*, vol. 8, pp. 82688–82699, 2020.

[44] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, Dec. 2018, pp. 7132–7141.

[45] I. Goodfellow, Y. Bengio, and A. Courville, " 6.2. 2.3 Softmax units for multinoulli output distributions," in *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016, pp. 180–184.

[46] J. Yerushalmy, "Statistical problems in assessing methods of medical diagnosis, with special reference to X-ray techniques," in *Proc. Conf. Health Rep.*, 1947, pp. 1432–1449.

[47] D. M. W. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," 2020, *arXiv:2010.16061*. [Online]. Available: http://arxiv.org/abs/2010.16061

[48] C. E. Metz, "Basic principles of ROC analysis," *Seminars Nucl. Med.*, vol. 8, no. 4, pp. 283–298, Oct. 1978.

[49] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[50] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1492–1500.

[51] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[52] 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.

[53] 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 Conf. Comput. Vis. Pattern Recognit.*, Oct. 2018, pp. 6450–6459.

[54] B. Zhou, A. Andonian, A. Oliva, and A. Torralba, "Temporal relational reasoning in videos," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 803–818.