Journal of Imaging Informatics in Medicine
https://doi.org/10.1007/s10278-024-00987-1

Atrial Septal Defect Detection in Children Based on Ultrasound Video Using Multiple Instances Learning

Yiman Liu[1,2,3] · Qiming Huang[4] · Xiaoxiang Han[5] · Tongtong Liang[6] · Zhifang Zhang[1] · Xiuli Lu[7] · Bin Dong[2] · Jiajun Yuan[2] · Yan Wang[3] · Menghan Hu[3] · Jinfeng Wang[4] · Angelos Stefanidis[4] · Jionglong Su[4] · Jiangang Chen[3] · Qingli Li[3] · Yuqi Zhang[1,2]

Received: 25 September 2023 / Revised: 9 December 2023 / Accepted: 12 December 2023

Abstract

Thoracic echocardiography (TTE) can provide sufficient cardiac structure information, evaluate hemodynamics and cardiac function, and is an effective method for atrial septal defect (ASD) examination. This paper aims to study a deep learning method based on cardiac ultrasound video to assist in ASD diagnosis. We chose four standard views in pediatric cardiac ultrasound to identify atrial septal defects; the four standard views were as follows: subcostal sagittal view of the atrium septum (subSAS), apical four-chamber view (A4C), the low parasternal four-chamber view (LPS4C), and parasternal short-axis view of large artery (PSAX). We enlist data from 300 children patients as part of a double-blind experiment for five-fold cross-validation to verify the performance of our model. In addition, data from 30 children patients (15 positives and 15 negatives) are collected for clinician testing and compared to our model test results (these 30 samples do not participate in model training). In our model, we present a block random selection, maximal agreement decision, and frame sampling strategy for training and testing respectively, resNet18 and r3D networks are used to extract the frame features and aggregate them to build a rich video-level representation. We validate our model using our private dataset by five cross-validation. For ASD detection, we achieve 89.33 ± 3.13 AUC, 84.95 ± 3.88 accuracy, 85.70 ± 4.91 sensitivity, 81.51 ± 8.15 specificity, and 81.99 ± 5.30 F1 score. The proposed model is a multiple instances learning-based deep learning model for video atrial septal defect detection which effectively improves ASD detection accuracy when compared to the performances of previous networks and clinical doctors.

Keywords Deep learning · Atrial septal defect · Multiple instances learning · Ultrasound video

Introduction

Congenital heart defects (CHDs) are the most common birth defects, with an incidence rate of approximately 0.9% in live births. They are the main causes of death in children aged between 0 and 5 years old [1]. The atrial septum separates the left and right atrium during the embryonic period. The abnormal formation of this gap may lead to CHDs after birth. If the gap does not close by itself during growth and development, atrial septal defect (ASD) occurs [2]. Based on the occurrence of ASD can be classified into two categories: primary ASD and secundum ASD. The secundum ASD can be divided into four types: central defect, superior cavity defect, inferior cavity defect, and mixed defect according to the location of the defect [3]. The type of defect studied in this paper was secundum ASD. ASD in children is one of the most common CHDs, accounting for approximately 6–10% of CHDs, with an estimated occurrence rate of 1–3 per 1000 [4]. The effect of ASD on the heart is a gradual process, starting from the right atrium volume overload to the right atrium expansion, and then to the right ventricle and pulmonary circulation system expansion, thus developing from the atrial shunt to pulmonary hypertension [2]. Most children with isolated atrial septal defects are free of symptoms, but the rates of exercise intolerance, arrhythmias, right ventricular failure, and pulmonary hypertension increase with age, and life expectancy is reduced in adults with untreated defects [5]. Therefore, accurate detection of ASD in early childhood is of great clinical importance.

Transthoracic echocardiography (TTE) can provide sufficient structural information about the heart and can be used

Yiman Liu, Qiming Huang, and Xiaoxiang Han contributed equally to this work.

Extended author information available on the last page of the article

Published online: 12 February 2024

to assess hemodynamics and cardiac function, serving as an effective method for ASD examination. The left and right atrial level shunt signals displayed by color Doppler flow imaging in echocardiography can accurately diagnose ASD [6]. The advantages of echocardiography include simplicity, non-invasiveness, low cost, repeatability, and accuracy. However, accurate diagnosis by TTE relies heavily on the accurate judgment of echocardiographers, which is time-consuming and subjective [7]. In China, especially in the regions with insufficient healthcare, there is a shortage of experienced cardiac ultrasound doctors. Meanwhile, poor-quality and noisy echocardiograms can lead to missed diagnoses and diagnostic errors as echocardiograms are affected by the machine imaging quality. It has been reported that up to 30% of ultrasound results are not very accurate [8]. Therefore, the research on the efficient and precise intelligent diagnostic of ASD is of paramount practical significance.

The development of deep learning (DL) [9, 10], was soon applied to the field of image analysis [11–13]. Moreover, there has been much research work on the analysis of ultrasound [14, 15] and echocardiography [16–19]. DL was recently shown to achieve exceptional performance in the realm of video understanding. The value of deep learning has been shown in 2D [20–23] and 3D networks [24–26]. The model we present in this paper is largely inspired by the aforementioned work in video understanding and offers improvement to it. There are some differences between the video understanding task and the video-based ASD recognition task that this article attempts to solve. Video understanding often involves providing a 3–5 s video and training the network to recognize its semantics. In such cases, the network pays more attention to the time-allowed information between video frames. In the case of video understanding, the model places more focus on the temporal information between video frames. However, during the ASD video recognition task, the video frames of most positive samples only appear for a short period of time. Therefore, how highlighting this segment is an issue that needs to be addressed when designing ASD recognition networks.

In this work, we propose a multi-instance learning-based method for the automatic recognition of ASD in cardiac ultrasound videos through attention aggregation across multiple video frames. Figure 3 illustrates the overview of the proposed approach. In a labeled as an ASD transthoracic echocardiograph video, only a few video frames may exhibit positive indicators of ASD, while the majority of the rest frames show negative ASD indicators. Utilizing multiple frames of the video directly for ASD recognition may lead to the omission of frames with positive indicators of ASD, introducing a potential bias in the prediction results. Hence, to mitigate these biases, we propose random block sampling and maximum agreement decision methods to use whole video frames in both the training and inference stages.

Materials and Methods

Participants

The subjects of this retrospective study were pediatric patients undergoing color Doppler echocardiography at Shanghai Children's Medical Center from March 1, 2022 to July 1, 2023. These cases included patients diagnosed with ASD (labeled as positive) and normal (as negative).

Data Collection

This study was approved by the Institutional Review Board of Shanghai Children's Medical Center (Approval No.SCMCIRB-K2022183-1) with a patient exemption applied. All patients were examined via echocardiography using GE Vivid E90 and E95 ultrasound systems with M5Sc and 6 S transducers(GE Healthcare is headquartered in Pudong New Area, Shanghai, China). Each echocardiographic study was conducted using standardized methods, as outlined by the American Society of Echocardiography (ASE) guidelines [27]. The four recommended standard views for atrial septal defect (ASD) detection are as follows: (1) subcostal sagittal view of the atrium septum (subSAS), (2) apical four-chamber view (A4C), (3) low parasternal short-axis four-chamber view (LPS4C), and (4) parasternal short-axis view of large artery (PSAX), and acquired their color Doppler dynamic ultrasound videos. All dynamic videos were blind reviewed by two senior sonographers and the videos with blurred atrial septum were excluded from the review. Standard imaging techniques were for two-dimensional, M-mode, and color Doppler echocardiography in accordance with the recommendations by the American Society of Echocardiography [28]. All data was strictly de-identified to protect patient privacy. The original data was stored in the Digital Imaging and Communications in Medicine (DICOM) format within the Picture Archiving and Communication Systems (PACS) database. To facilitate program processing, DICOM data is converted into dynamic AVI video.

Training/Validation Dataset

A total of 300 patients (comprising 1057 dynamic Doppler ultrasound videos, the distribution of the number of videos contained in each standard view is presented in Table 1.) are used as the training and validation set for ASD detection, as shown in Fig. 1. The visualization of typical ASD and normal static images from the collected dataset is given in Fig. 2. The recruitment of patients is initialized from the Ultrasound Report System of The Department of Pediatric Cardiology, the hospital by retrospective searching. The

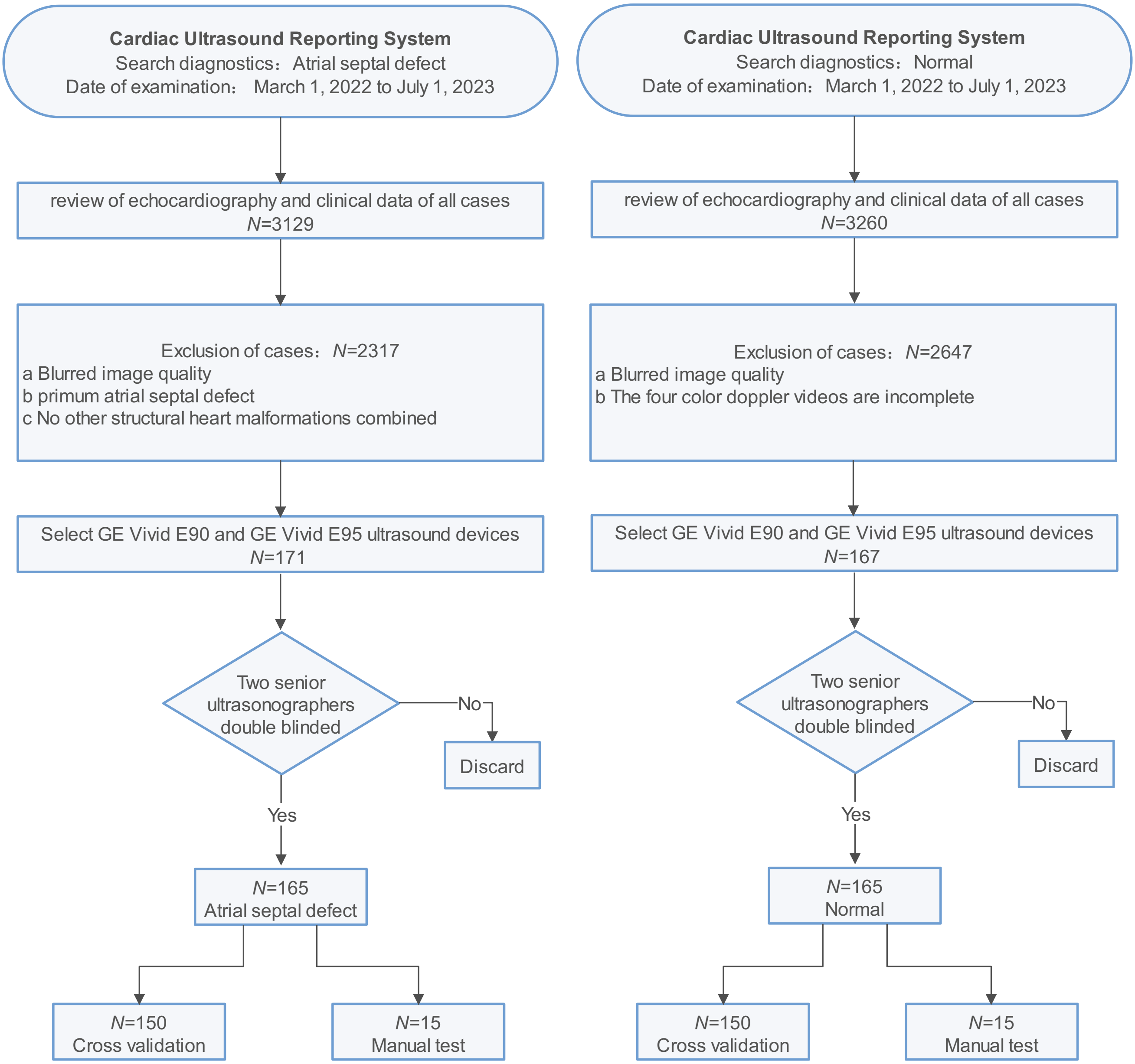
Cardiac Ultrasound Reporting System
Search diagnostics：Atrial septal defect
Date of examination：March 1, 2022 to July 1, 2023
review of echocardiography and clinical data of all cases
N=3129
Exclusion of cases：N=2317
a Blurred image quality
b primum atrial septal defect
c No other structural heart malformations combined
Select GE Vivid E90 and GE Vivid E95 ultrasound devices
N=171
Two senior ultrasonographers double blinded
No
Discard
Yes
N=165
Atrial septal defect
N=150
Cross validation
N=15
Manual test
Cardiac Ultrasound Reporting System
Search diagnostics：Normal
Date of examination：March 1, 2022 to July 1, 2023
review of echocardiography and clinical data of all cases
N=3260
Exclusion of cases：N=2647
a Blurred image quality
b The four color doppler videos are incomplete
Select GE Vivid E90 and GE Vivid E95 ultrasound devices
N=167
Two senior ultrasonographers double blinded
No
Discard
Yes
N=165
Normal
N=150
Cross validation
N=15
Manual test

Fig. 1 Inclusion and exclusion criteria for the main cohort of this study

diagnostic keyword used is set as the “Atrial Septal Defect,” and the time of examination is set from March 1, 2022 to July 1, 2023. The obtained cases are reviewed consecutively to evaluate candidacy based on echocardiography results and clinical data. The process is given in Fig. 2. The final number of cases in ASD and normal are 165 and 165 respectively. While 150 cases in ASD and 150 cases in normal for cross-validation as well as 15 cases in ASD and 15 cases in normal for manual testing.

Specifically, the inclusion criteria for the data collection are as follows: (1) secundum atrial septal defect and (2) the use of GE Vivid E90 and GE Vivid E95 ultrasound devices.

A few exclusion criteria are also in place to avoid interference with the training results of the model: (1) blurred image quality; (2) primum atrial septal defect; and (3) no other structural heart malformations combined. The demographic and clinical characteristics of the training and validation data sets are given in Table 2.

Methods

In this study, we propose an echocardiography video-based atrial septal defect diagnosis system. The overall model structure is given in Fig. 3. We propose a block random frame

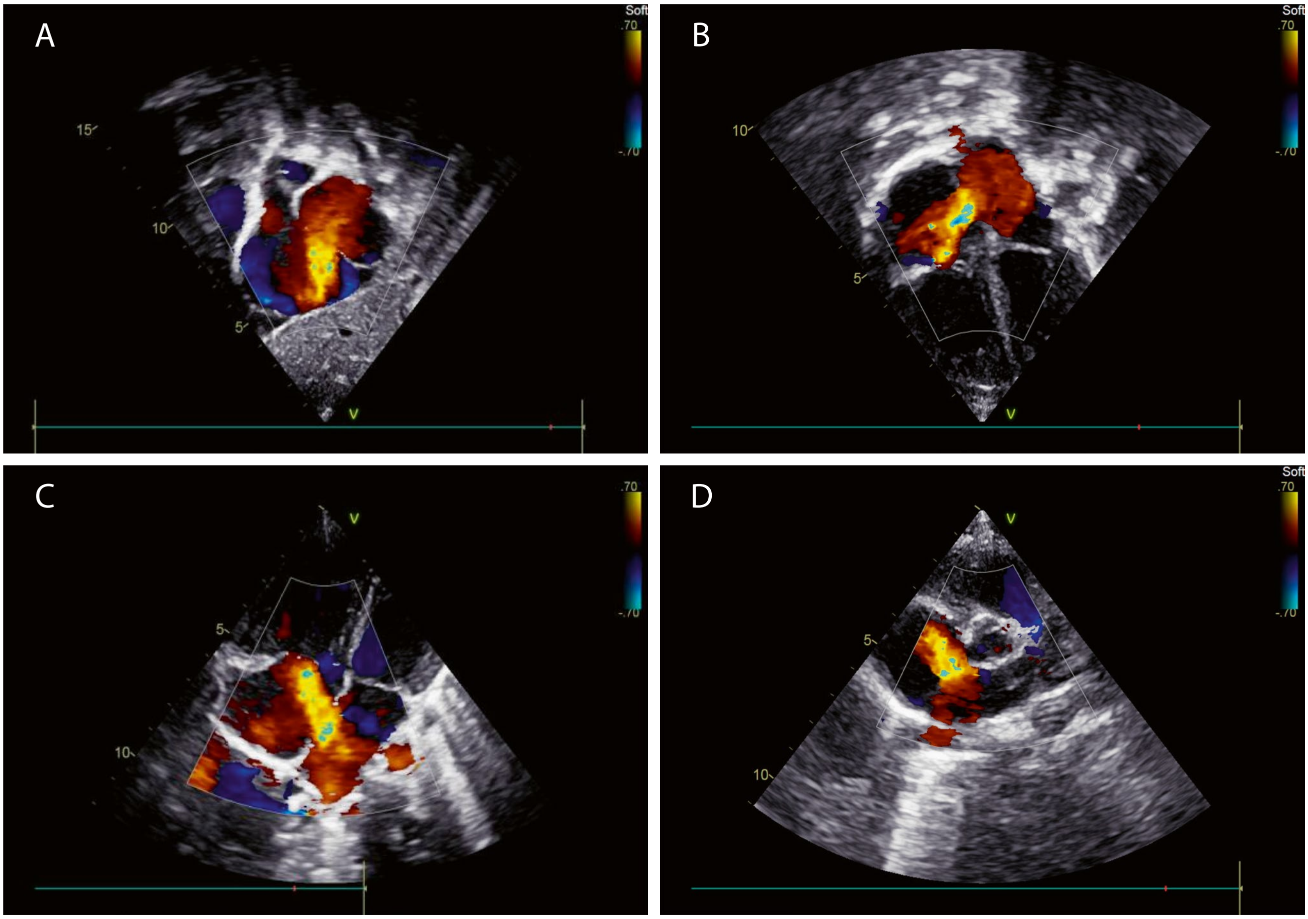

Fig. 2 Visualization of positive samples from four standard sections of ASD. **A** Subcostal sagittal view of the atrium septum (subSAS). **B** Apical four-chamber view (A4C). **C** Low parasternal four-chamber view (LPS4C). **D** Parasternal short-axis view of large artery (PSAX)

sampling strategy for the training stage to train our model robustly. The maximum agreement decision during the testing stage is offered to refine inference prediction. Then, resNet18 is used as a feature extractor to capture the frame features. To enrich the temporal information between different echocardiography frames, we use 3D convolution after stage 2 in resNet18 and generate temporal representation. Afterward, for each frame feature, attention pooling is used to aggregate them to form a spatial representation. Finally, we combine the spatial representation and temporal representation to build a general video-level representation for ASD classification.

Multiple Instance Learning by Attention Aggregation

The video-based echocardiography ASD detection problem can be considered as a multiple-instance learning problem.

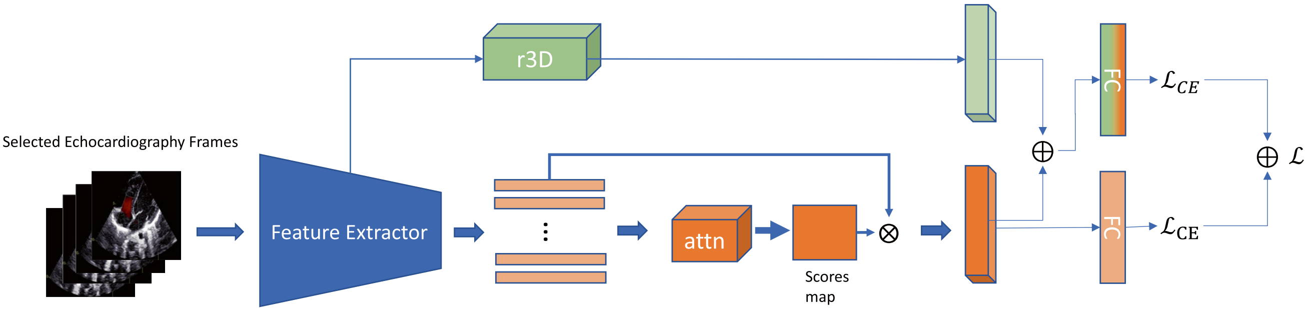

Fig. 3 The architecture of our proposed model for ASD detection. It contains two branches. One branch is employed to extract the spatial features of the video, while the other branch employs an r3D network to capture the temporal features of the video. These two features are then concatenated for final classification

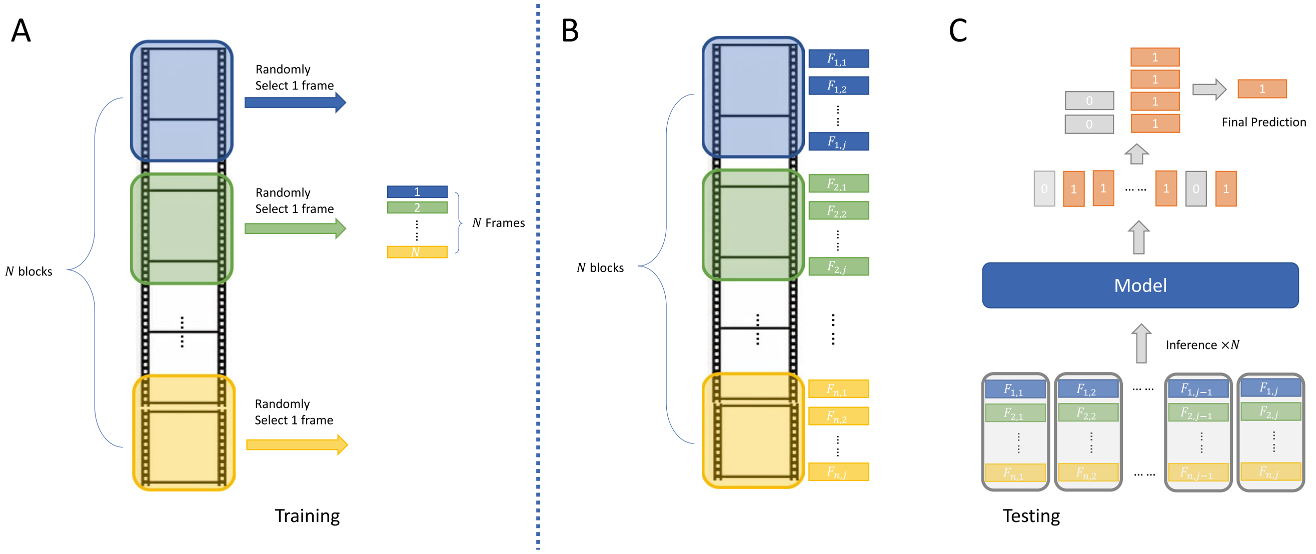

Fig. 4 The proposed frame selection strategies on training and testing stages. **A** Block random selection (BRS) for training. **B** Block inference for testing. We assign each echocardiography frame with a frame index $F_{i,j}$ where i, j indicates the j^{th} frame in i^{th} block. **C** Our proposed decision rule called maximum agreement decision (MAD) to refine ASD prediction

Let B_i denote the i^{th} bag of frames in echocardiography videos. Let B_{ij} denote the j^{th} frames in bag $B_i, j = 1, \ldots, n_i$ with a total of n_i frame. Let $\mathcal{F}_{B_{ij}}$ denote the frame feature given by $F(B_{ij})$ where $F(\cdot)$ is deep feature extractor. Let $\mathcal{F}_{B_i}$ denote the video feature given by $\mathcal{G}(\mathcal{F}(B_{ij}))$ where $\mathcal{G}$ is a feature aggregation function. Finally, let y_i denote the label of bag B_i, where $y_i \in \{-1, 1\}$. The aim is to establish a linear mapping function with projection $\mathcal{F}_{B_i} \rightarrow y_i$. In the realm of deep learning, this can be expressed as a fully connected layer, mapping the D-dimensional features of an image or video to the dimensionality of the classes. In the work, the prediction class dimension is 2, representing ASD or normal.

In ASD video recognition, each video frame segment can be regarded as an instance, and each video is considered a bag. The individual labels for each video frame segment are unknown, we only have access to the labels for the entire video annotated by clinical experts. In this scenario, the selection of video frames for model learning is a crucial consideration. The most intuitive approach is to extract features from each frame of the video and then integrate these features to form a comprehensive representation. Two common aggregation methods are $average(\cdot)$ and $max(\cdot)$ where $average(\cdot)$ takes the mean of all frame features to represent a video feature and $max(\cdot)$ takes maximum frame feature to represent a video feature. However, in the case of the average approach, it assigns equal weight to each instance, overlooking the varying significance of positive and negative samples in ASD video recognition. Meanwhile, the maximum-value method utilizes the features of only one frame to represent the entire video, which may not capture the full context comprehensively. Hence, we instead use the attention aggregation module proposed in [29] to aggregate the features. Formally, we let $\mathcal{H} = \{h_1, h_2, \ldots, h_J\}$ denote a video with J frames. The attention aggregation function is given by the following:

$$\mathcal{Z} = \sum_{j=1}^{J} a_j h_j \tag{1}$$

where

$$\alpha_j = \frac{\exp\{\boldsymbol{w}^T \tanh(\boldsymbol{V}\boldsymbol{h}_j^T)\}}{\sum_{j=1}^{J} \exp\{\boldsymbol{w}^T \tanh(\boldsymbol{V}\boldsymbol{h}_j^T)\}} \tag{2}$$

$tanh(\cdot)$ is a non-linearity activation function. M is the dimension of the mapped feature. We set $M = 1024$ instead of the default value of 512 in [29] to strengthen

Table 1 Distribution of number of videos in ASD and normal groups

Standard view	ASD group (Number of videos)	Normal group (Number of videos)	Video time (second)	Cardiac cycle (number)
subSAS	150	150	3	3
A4C	81	150	3	3
LPS4C	150	150	3	3
PSAX	76	150	3	3

Table 2 Clinical characteristic comparisons between the ASD group and the normal group of the data set where n denotes the number of cases

Characteristics	**ASD group (n = 150)**	**Normal group (n = 150)**	**p -value**
Age (years)	2.66(1.41−4.68)	2.23(0.91−5.23)	0.15
Female/male	88/62	82/68	0.49
Weight (kg)	13.50(10.08−17.58)	11.10(7.60−18.53)	0.06
Height (cm)	80.00(93.00−110.00)	87.50(69.75−113.25)	0.05
Size of ASD (mm)	0.64(0.94−1.53)	-	-

the capacity to model complicated data. By employing attention aggregation, we can dynamically assign weights to each instance based on the similarity between video frames. Ultimately, the feature representation of the video is a weighted average across all instances, reflecting their individual contributions.

As stated earlier, there are typically two network architectures for video classification: 2D-based and 3D-based. For example, 3D convolutional networks [24, 30, 26] are widely used in video understanding. 3D convolution is used to model the temporal information. However, we argue that in video-based atrial septal defect detection, spatial information is not the principal factor, since a positive frame may appear within a short period of time. Hence, we build our classification model based on the 2D convolutional network. The overall model structure is depicted in Fig. 3. For each frame, we use a resNet18 [9] to obtain the feature representation and concatenate them to form an $N \times D$ general video representation where N is the number of frames used for training and D is the dimension of the feature. Following this, we use attention pooling to generate $N \times 1$ weights and multiply them by the general feature representation to form the final spatial representation $\mathcal{D}_{spatial}$. Although temporal information is not dominant in ASD detection, we can use 3D convolution [31] to model the temporal information of color Doppler signals. Similar to [32], the middle features from $conv2_x, conv3_x, conv4_x$ of r3d (detail of r3d is given in Table 3) are extracted to capture temporal information between different training frames and generate a temporal feature representation $\mathcal{D}_{temporal}$. The final representation is given by the following:

$$\mathcal{D} = \mathcal{D}_{spatial} \oplus \mathcal{D}_{temporal} \tag{3}$$

where ⊕ denotes the element-wise addition.

Table 3 The architectures of r3d [26]. For each layer, it repeats 4 times. The sequence of convolutional operations concludes with a global spatiotemporal pooling layer, producing a 512-dimensional feature vector. This vector is then input into a fully connected layer, which generates class probabilities using a softmax activation

layer name	output size	r3d
conv1	$L \times 56 \times 56$	$3 \times 7 \times 7, 64, \text{stride}1 \times 2 \times 2$
conv2_x	$L \times 56 \times 56$	$3 \times 3 \times 3, 64$
conv3_x	$\frac{L}{2} \times 28 \times 28$	$3 \times 3 \times 3, 128$
conv4_x	$\frac{L}{4} \times 14 \times 14$	$3 \times 3 \times 3, 256$
conv5_x	$\frac{L}{8} \times 7 \times 7$	$3 \times 3 \times 3, 512$
	$1 \times 1 \times 1$	spatiotemporal pooling, fc layer with softmax

Frame Sampling Strategy and Decision Rule

Since the video frames are computationally more expensive, using all frames in the training stage is time-consuming and inefficient. Furthermore, adjacent video clips have high similarities which may lead to overfitting. To address these, we propose two different frame selection strategies for the training stage and inference stage respectively. Figure 4 gives the overall frame sampling method. Similar to [21], we segment each video into N blocks with each block containing the same number of frames. Subsequently, we randomly select one frame from the whole video to construct an N-frames collection used for training. This can be regarded as data augmentation used to improve the robustness of the model. Compared to randomly sampling N frames from the whole video, this method ensures the diversity of frames since we strictly divide the whole video into several blocks and sample frames from each of these blocks.

As a positive sample of atrial septal defect can manifest in any frame within the video, selecting only certain frames may result in misclassification. Therefore, we employ all frames for inference and introduce a decision rule known as maximal agreement decision, inspired by the majority rule observed in human society. Specifically, in the inference stage, we also split the whole video into N blocks, for each block, we select i frame, $i \in 1, ..., K$, where K is the number of frames in a video block. $B_i, i \in 1, ..., N$ is denoted as the i th frames collection used for testing. We then have N predictions $[Y_1, ..., Y_N], Y_i \in \{-1, +1\}, \forall i \in \{1, 2, ..., n\}$. The final prediction $\hat{Y}$ for this video is determined by the following:

$$\hat{Y} = \max\left\{\sum_{i=1}^{N} Y_i = -1, \sum_{i=1}^{N} Y_i = 1\right\} \tag{4}$$

Table 4 Clinical characteristic comparisons between the ASD group and the normal group of the test data set where n denotes the number of cases. *The overall data of ASD (atrial septal defect) size did not conform to normal distribution, and the difference in ASD size between the training set and the test set was statistically tested by non-parametric test, $P = 0.301$, which could not reject the null hypothesis at the test level of 0.05, therefore, there was no statistically significant difference in ASD size between the training set and the test set

Characteristics	**ASD group ($n = 15$)**	**Normal group ($n = 15$)**	***p*** **-value**
Age (years)	3.33(1.33−5.74)	3.57(1.25−5.58)	0.96
Female/male	9/6	8/7	0.72
Weight (kg)	16.00(11.50−17.50)	16.50(10.50−18.00)	0.74
Height (cm)	85.00(105.00−110.00)	86.00(106.00−112.00)	0.86
Size of ASD (mm)	0.65(1.22−1.93)*	-	-

This decision rule mimics the difficult and severe diseases in real clinical practice. We consider Y_i as conclusions from all experts. Then, the major rule is used to obtain the final diagnosis.

Experiments

Experiments Setup

We implement our model in PyTorch using an NVIDIA TESLA T4 GPU accelerator. The SGD optimizer is used with an initial learning rate of 1e-4. During training, we resize the video frames to 224×224, we select $N = 16$ frames from videos for training with batch size 32. In the main experiment, five-fold cross-validation is used to evaluate the performance of the proposed model. There is no data augmentation used in training.

Three hundred and thirty cases are used in this study. Specifically, three hundred cases are used to assess the performance of the proposed method. The remaining 30 cases are excluded from the training and are directly used for additional testing and their demographic and clinical characteristics of the dataset are shown in Table 4.

Evaluation Protocols

We mainly evaluate our proposed model using confusion matrix analysis, described by the accuracy, sensitivity, specificity, and F1 score.

Results and Discussion

Main Results

As shown in Table 5, compared to different models including 3D CNN-based, our proposed model achieves the best performance with 89.33 ± 3.13 AUC and 84.95 ± 3.88 accuracy on 5-fold cross-validation. The details of sensitivity, specificity, and F1 score are given in Table 5. In addition, we compare the ASD classification accuracy with doctors‘ diagnoses of ASD. In 30 cases of testing both junior and senior doctors can achieve ASD detection accuracy of 63.66% and 71.70% respectively, while our model achieves 83.33% accuracy as shown in Table 8.

Ablation Study

We also report the ablation study of our components, i.e., 3D fusion, attention aggregation, maximum agreement decision, and block random selection. The results, given in Tables 6 and 7, demonstrate the effectiveness of our proposed model for improving ASD detection accuracy. Furthermore, we provide an ablation study pertaining to the impact of varying the number of frames selected for both training and testing, as detailed in Table 9. As we hope the model can be applied in clinical auxiliary diagnosis, our emphasis lies in ensuring its compatibility with lightweight deployment requirements. The comparative analysis presented in Table 10, it exhibits the lowest parameter count and computational cost.

Table 5 The ASD classification result of different models via five five-fold cross-validations (Mean±Std), *The AUC serves as a comprehensive metric reflecting the magnitude of diagnostic test value. In this study, ResNet18 was utilized as a reference model for comparative analysis. The statistical differences in AUC between ResNet18 and three other models, namely X3D, R2plus1D, and Ours, were examined. The results of a one-sided t-test revealed statistically significant differences in AUC between ResNet18 and the aforementioned models

Model	AUC(%)	Accuracy(%)	Sensitivity (%)	Specificity(%)	F1(%)	*p*-value*
X3D [30]	72.12 ± 1.25	69.02 ± 2.98	72.69 ± 3.13	68.45 ± 4.12	71.72 ± 2.22	<0.0005
R2plus1D [26]	76.21 ± 1.45	74.09 ± 1.88	77.92 ± 1.07	73.34 ± 1.66	78.05 ± 1.80	<0.0005
ResNet18 [9]	85.25 ± 3.91	77.25 ± 3.92	78.19 ± 3.12	78.92 ± 2.92	76.87 ± 1.91	Ref
Ours	**89.56 ± 2.19**	**83.95 ± 4.12**	**84.10 ± 3.20**	**82.51 ± 8.04**	**82.99 ± 5.30**	<0.005

Table 6 Ablation study for 3D fusion, attention aggregation module (AAM) via five five-fold cross-validations (Mean±Std)

Model	3D fusion	AAM	Accuracy(%)
ResNet18	✗	✗	77.25 ± 3.82
	✗	✓	82.59 ± 2.83
	✓	✗	80.21 ± 2.11
	✓	✓	84.06 ± 3.38

Table 7 Ablation study for MAD, BRS via five five-fold cross-validations (Mean±Std) and our model refers to Resnet18+3D fusion+AAM

Model	MAD	BRS	Accuracy(%)
Our model	✗	✗	84.06 ± 3.38
	✗	✓	84.59 ± 2.83
	✓	✗	84.44 ± 4.41
	✓	✓	84.95 ± 3.88

Table 8 ASD detection performance of our model and professional doctors based on additional testing. *We compared the *p*-values between junior doctors and intermediate doctors as well as the model by using a statistical analysis using paired chi-square test with Accuracy of junior doctors as the reference

Model	Accuracy(%)	PPV(%)	NPV(%)	*p*-value*
Junior doctor	63.33	53.49	89.47	Ref
Senior doctor	71.70	60.53	91.67	0.359
Ours	83.33	85.71	81.25	0.029

Table 9 Accuracy results about different numbers of video frames selected during the training and testing

Number of frames	Accuracy
8	75.21
16	83.95
32	82.18

Table 10 The comparison between model parameters and computational cost

Model	Params (M)	MACs(G)
resnet18	11.60	1.82
x3d	34.21	216.30
r2plus1d	31.52	163.10
ours	12.10	6.36

Discussion

In Table 8, it can be seen that the accuracy for ASD detection of our proposed model is higher than that of both junior doctors and senior doctors detection rates. However, the NPV of our model is lower than that of the doctors. A similar result regarding early Alzheimer's disease prediction using deep learning is reported in [33]. Positive Predictive Value (PPV) is an indicator to measure the ability of the model to avoid misdiagnosis and Negative Predictive Value (NPV) is used to measure the missed diagnosis ability. Higher PPV can indicate the potential of our model to be used as an ASD assistance tool in clinical.

Furthermore, our introduced maximum agreement decision approach utilizes the entirety of video frames for video predictions, thereby mitigating the circumstance in which affirmative video frames are overlooked, consequently averting inaccuracy prediction. The attention module we have introduced exhibits the capability to dynamically amalgamate information based on feature similarity among video frames during the training phase. It effectively accentuates positive instances within numerous negative video frames by assigning them elevated weights. Our ablation experiments further substantiate that the incorporation of this attention mechanism has yielded a noteworthy performance enhancement of 5% as given in Table 6. The heat map of our model is given in Fig. 5. The most activated areas are shown by the color Doppler signal area which indicates the shunt blood flow signal. The importance of the color Doppler signal in the diagnosis of ASD is emphasized by this finding.

In addition, we have observed that ASD shunting of blood flow is not present in every frame of the cardiac cycle, owing to the systole and diastole of the heartbeat. To the best of our knowledge, this study attempts to diagnose ASD in children at the level of ultrasound videos. It has been reported that deep learning algorithms can assist inexperienced diagnosticians in obtaining echocardiograms for limited diagnoses [34]. Therefore, our ultrasound video diagnosis algorithm for ASD in children holds significant clinical importance: for experienced doctors, it can enhance the efficiency of film reviews and ensure the reliability of ASD diagnoses; for less experienced doctors or those lacking the ability to review films, it can provide echocardiography-based ASD diagnostic information, thus enhancing the comprehensiveness of the diagnosis.

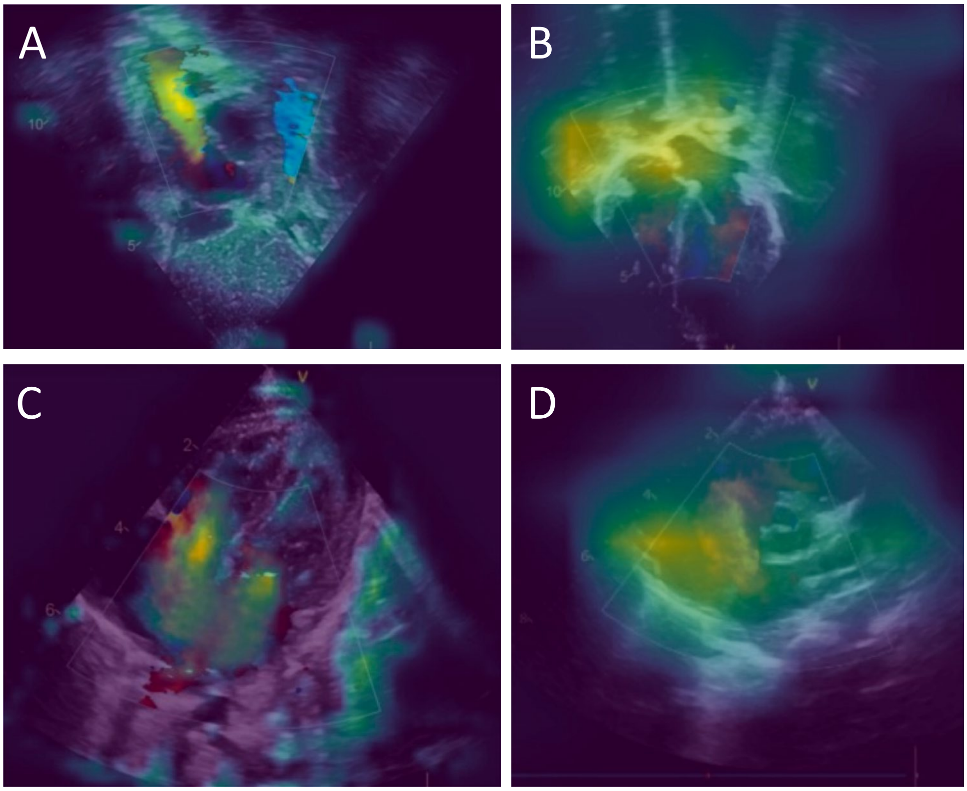

Fig. 5 Visualization of attention maps for four standard sections of ASD. **A** Subcostal view of the atrial septum (subAS) **B** Apical 4-chamber view(A4C). **C** The low parasternal four-chamber view (LPS4C). **D** Parasternal short-axis view of large artery (PSAX)

Conclusion

This paper proposes a multi-instance learning-based model for the diagnosis of atrial septal defects using ultrasound video that has achieved excellent results. Since atrial septal defect is a very common cardiac disease that is found in a large number of children, our method can not only help cardiac sonographers to improve the efficiency of diagnosis but also assist primary cardiac sonographers in making diagnostic decisions.

However, the limitations of our study are four-fold: First, we select ultrasound videos from using only four standard views for diagnosing atrial septal defects, while there are some other ultrasound videos from non-standard views that can diagnose atrial septal defects that are not included in our study; Second, our data sets were all secondary foramen ovale atrial septal defects and exclude primary foramen ovale atrial septal defects. In addition, the shunt direction of the defects is all left-to-right shunts from the left atrium to the right atrium; Finally, the amount of data in our article was small and single-centred. These aspects will be examined in our future work.

Author Contributions Yiman Liu: Study. Qiming Huang: Study. Xiaoxiang Han: Study. Tongtong Liang: Statistical analysis of data. Zhifang Zhang: Data collection. Xiuli Lu: Data collection. Bin Dong: Administrative support. Jiajun Yuan: Administrative support. Yan Wang: Model method guidance. Menghan Hu: Data collection, Analysis, Graphics. Jinfeng Wang: Data collection, Analysis, Graphics. Angelos Stefanidis: Model method guidance. Jionglong Su: Provided critical feedback and guidance throughout the project. Jiangang Chen: Provided critical feedback and guidance throughout the project. Qingli Li: Provided critical feedback and guidance throughout the project. Yuqi Zhang: Provided critical feedback and guidance throughout the project.

Funding This work is supported by the National Natural Science Foundation of China (Grant No. 61975056), the Shanghai Natural Science Foundation (Grant No. 19ZR1416000), the Science and Technology Commission of Shanghai Municipality (Grant No. 20440713100) and the Key Program State Fund in XJTLU (KSF-A-22), the Shanghai Municipal Health Commission (Project No. 20194Y0321), the Sanya Science and Technology Innovation Program(project No 2022KJCX41), Pediatric medical consortium scientific research project of Shanghai Children’s Medical Center affiliated to Shanghai Jiao Tong University School of Medicine.

Declarations

Ethics Approval This study was approved by the Institutional Review Board of Shanghai Children’s Medical Center (Approval No.SCMCIRB-K2022183-1) with a patient exemption applied.

Consent to Participate Informed consent was waived by ethics groups.

Consent to Publish Informed consent was waived by ethics groups.

Conflict of Interest The authors declare no competing interests.

References

1. Zhao, Q.-M., Liu, F., Wu, L., Ma, X.-J., Niu, C., Huang, G.-Y.: Prevalence of congenital heart disease at live birth in china. J Pediatr **204**, 53–58 (2019). https://doi.org/10.1016/j.jpeds.2018.08.040
2. Chen, H., Yan, S., Xie, M., Ye, Y., Ye, Y., Zhu, D., Su, L., Huang, J.: Fully connected network with multi-scale dilation convolution module in evaluating atrial septal defect based on mri segmentation. Comput Methods Programs Biomed **215**, 106608 (2022). https://doi.org/10.1016/j.cmpb.2021.106608
3. Rhodes, J., Patel, H., Hijazi, Z.M.: Effect of transcatheter closure of atrial septal defect on the cardiopulmonary response to exercise. Am J Cardiol **90**(7), 803–806 (2002). https://doi.org/10.1016/s0002-9149(02)02620-6
4. Bradley, E.A., Zaidi, A.N.: Atrial septal defect. Cardiol Clin **38**(3), 317–324 (2020). https://doi.org/10.1016/j.ccl.2020.04.001
5. Geva, T., Martins, J.D., Wald, R.M.: Atrial septal defects. Lancet **383**(9932), 1921–1932 (2014). https://doi.org/10.1016/S0140-6736(13)62145-5
6. Huang, S., Liu, J., Lee, L.C., Venkatesh, S.K., Teo, L.L.S., Au, C., Nowinski, W.L.: An image-based comprehensive approach for automatic segmentation of left ventricle from cardiac short axis cine mr images. J Digit Imaging **24**(4), 598–608 (2011). https://doi.org/10.1007/s10278-010-9315-4
7. Wu, L., Dong, B., Liu, X., Hong, W., Chen, L., Gao, K., Sheng, Q., Yu, Y., Zhao, L., Zhang, Y.: Standard echocardiographic view recognition in diagnosis of congenital heart defects in children using deep learning based on knowledge distillation. Front Pediatr **9** (2021). https://doi.org/10.3389/fped.2021.770182
8. Lu, Y., Radau, P., Connelly, K., Dick, A., Wright, G.A.: Segmentation of left ventricle in cardiac cine mri: An automatic image-driven method. In: International Conference on Functional Imaging and Modeling of the Heart, pp. 339–347 (2009). https://doi.org/10.1007/978-3-642-01932-6_37. Springer
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016). https://doi.org/10.48550/arXiv.1512.03385
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., *et al*: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020). https://doi.org/10.48550/arXiv.2010.11929
11. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, pp. 234–241 (2015). https://doi.org/10.1007/978-3-319-24574-4_28. Springer
12. Yadav, D., Jain, R., Agrawal, H., Chattopadhyay, P., Singh, T., Jain, A., Singh, S.B., Lee, S., Batra, D.: Evalai: Towards better evaluation systems for ai agents. arXiv preprint arXiv:1902.03570 (2019). https://doi.org/10.48550/arXiv.1902.03570
13. Sekuboyina, A., Husseini, M.E., Bayat, A., Löffler, M., Liebl, H., Li, H., Tetteh, G., Kukačka, J., Payer, C., Štern, D., *et al*: Verse: A vertebrae labelling and segmentation benchmark for multi-detector ct images. Med Image Anal **73**, 102166 (2021). https://doi.org/10.1016/j.media.2021.102166
14. Lin, Z., Lin, J., Zhu, L., Fu, H., Qin, J., Wang, L.: A new dataset and a baseline model for breast lesion detection in ultrasound videos. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part III, pp. 614–623 (2022). https://doi.org/10.1007/978-3-031-16437-8_59. Springer
15. Chen, Y., Zhang, C., Liu, L., Feng, C., Dong, C., Luo, Y., Wan, X.: Uscl: pretraining deep ultrasound image diagnosis model through video contrastive representation learning. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24, pp. 627–637 (2021). https://doi.org/10.1007/978-3-030-87237-3_60. Springer
16. Huang, Z., Long, G., Wessler, B., Hughes, M.C.: A new semi-supervised learning benchmark for classifying view and diagnosing aortic stenosis from echocardiograms. In: Machine Learning for Healthcare Conference, pp. 614–647 (2021). https://doi.org/10.48550/arXiv.2108.00080. PMLR
17. Nagueh, S.F.: Left ventricular diastolic function: understanding pathophysiology, diagnosis, and prognosis with echocardiography. JACC Cardiovasc Imaging **13**(1 Part 2), 228–244 (2020). https://doi.org/10.1016/j.jcmg.2018.10.038
18. Østvik, A., Salte, I.M., Smistad, E., Nguyen, T.M., Melichova, D., Brunvand, H., Haugaa, K., Edvardsen, T., Grenne, B., Lovstakken, L.: Myocardial function imaging in echocardiography using deep learning. IEEE Trans Med Imaging **40**(5), 1340–1351 (2021). https://doi.org/10.1109/TMI.2021.3054566
19. Ahn, S.S., Ta, K., Thorn, S.L., Onofrey, J.A., Melvinsdottir, I.H., Lee, S., Langdon, J., Sinusas, A.J., Duncan, J.S.: Co-attention spatial transformer network for unsupervised motion tracking and cardiac strain analysis in 3d echocardiography. Med Image Anal **84**, 102711 (2023). https://doi.org/10.1016/j.media.2022.102711
20. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. Advances in neural information processing systems **27** (2014). https://doi.org/10.48550/arXiv.1406.2199
21. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks: Towards good practices for deep action recognition. In: European Conference on Computer Vision, pp. 20–36 (2016). https://doi.org/10.1007/978-3-319-46484-8_2. Springer
22. Lin, R., Xiao, J., Fan, J.: Nextvlad: An efficient neural network to aggregate frame-level features for large-scale video classification. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops, pp. 0–0 (2018). https://doi.org/10.1007/978-3-030-11018-5_19
23. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6202–6211 (2019). https://doi.org/10.1109/ICCV.2019.00630
24. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4489–4497 (2015). https://doi.org/10.1109/ICCV.2015.510
25. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6299–6308 (2017). https://doi.org/10.1109/CVPR.2017.502
26. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6450–6459 (2018). https://doi.org/10.1109/CVPR.2018.00675
27. Silvestry, F.E., Cohen, M.S., Armsby, L.B., Burkule, N.J., Fleishman, C.E., Hijazi, Z.M., Lang, R.M., Rome, J.J., Wang, Y.: Guidelines for the echocardiographic assessment of atrial septal defect and patent foramen ovale: from the american society of echocardiography and society for cardiac angiography and interventions. Journal of the American Society of Echocardiography **28**(8), 910–958 (2015). https://doi.org/10.1016/j.echo.2015.05.015

28. Lopez, L., Colan, S.D., Frommelt, P.C., Ensing, G.J., Kendall, K., Younoszai, A.K., Lai, W.W., Geva, T.: Recommendations for quantification methods during the performance of a pediatric echocardiogram: a report from the pediatric measurements writing group of the american society of echocardiography pediatric and congenital heart disease council. J Am Soc Echocardiogr **23**(5), 465–495 (2010). https://doi.org/10.1016/j.echo.2010.03.019
29. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International Conference on Machine Learning, pp. 2127–2136 (2018). https://doi.org/10.48550/arXiv.1802.04712. PMLR
30. Feichtenhofer, C.: X3d: Expanding architectures for efficient video recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 203–213 (2020). https://doi.org/10.1109/CVPR42600.2020.00028
31. Hara, K., Kataoka, H., Satoh, Y.: Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6546–6555 (2018). https://doi.org/10.1109/CVPR.2018.00685
32. Zolfaghari, M., Singh, K., Brox, T.: Eco: Efficient convolutional network for online video understanding. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 695–712 (2018). https://doi.org/10.1007/978-3-030-01216-8_43
33. Hu, Z., Wang, Z., Jin, Y., Hou, W.: Vgg-tswinformer: Transformer-based deep learning model for early alzheimer's disease prediction. Comput Methods Programs Biomed **229**, 107291 (2023). https://doi.org/10.1016/j.cmpb.2022.107291
34. Narang, A., Bae, R., Hong, H., Thomas, Y., Surette, S., Cadieu, C., Chaudhry, A., Martin, R.P., McCarthy, P.M., Rubenson, D.S., *et al*: Utility of a deep-learning algorithm to guide novices to acquire echocardiograms for limited diagnostic use. JAMA Cardiol **6**(6), 624–632 (2021). https://doi.org/10.1001/jamacardio.2021.0185

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Yiman Liu[1,2,3] · Qiming Huang[4] · Xiaoxiang Han[5] · Tongtong Liang[6] · Zhifang Zhang[1] · Xiuli Lu[7] · Bin Dong[2] · Jiajun Yuan[2] · Yan Wang[3] · Menghan Hu[3] · Jinfeng Wang[4] · Angelos Stefanidis[4] · Jionglong Su[4] · Jiangang Chen[3] · Qingli Li[3] · Yuqi Zhang[1,2]

✉ Jionglong Su
Jionglong.Su@xjtlu.edu.cn

✉ Jiangang Chen
jgchen@cee.ecnu.edu.cn

✉ Qingli Li
qlli@cs.ecnu.edu.cn

✉ Yuqi Zhang
changyuqi@hotmail.com

Yiman Liu
LiuyimanSCMC@163.com

Qiming Huang
Qiming.Huang19@student.xjtlu.edu.cn

Xiaoxiang Han
gtlinyer@163.com

Tongtong Liang
zhimaliang@163.com

Zhifang Zhang
zzftcy@126.com

Xiuli Lu
529778564@qq.com

Bin Dong
dongbin@scmc.com.cn

Jiajun Yuan
yuanjiajun@scmc.com.cn

Yan Wang
ywang@cee.ecnu.edu.cn

Menghan Hu
mhhu@ce.ecnu.edu.cn

Jinfeng Wang
Jinfeng.Wang20@student.xjtlu.edu.cn

Angelos Stefanidis
Angelos.Stefanidis@xjtlu.edu.cn

[1] Department of Pediatric Cardiology, Shanghai Children's Medical Center, School of Medicine, Shanghai Jiao Tong University, Shanghai 200127, China

[2] Shanghai Engineering Research Center of Intelligence Pediatrics (SERCIP), Shanghai 200127, People's Republic of China

[3] Shanghai Key Laboratory of Multidimensional Information Processing, school of communication and electronic engineering, East China Normal University, Shanghai 200241, People's Republic of China

[4] School of AI and Advanced Computing, Xi'an Jiao tong-Liverpool University, Taicang 215028, People's Republic of China

[5] School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai 200093, People's Republic of China

[6] Shanghai Minhang Center for Disease Control and Prevention, Shanghai 201101, People's Republic of China

[7] Department of Ultrasound, Jiaxing Xiuzhou District Maternal, Child Health Hospital, Jiaxing, Zhejiang 314031, People's Republic of China