



Contents lists available at ScienceDirect

Medical Image Analysis

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

Automated interpretation of congenital heart disease from multi-view echocardiograms

Jing Wang^{a,1}, Xiaofeng Liu^{c,f,1}, Fangyun Wang^b, Lin Zheng^b, Fengqiao Gao^a, Hanwen Zhang^a, Xin Zhang^b, Wanqing Xie^d, Binbin Wang^e^aDepartment of Medical Genetics and Developmental Biology, School of Basic Medical Sciences, Capital Medical University, Beijing, 10069, China^bHeart Center, Beijing Children's Hospital, Capital Medical University, National Center for Children's Health, Beijing, 10045, China^cDepartment of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA, 15232 USA^dCollege of Mathematical Sciences, Harbin Engineering University, Harbin, China^eCenter for Genetics, National Research Institute for Family Planning, Beijing, China^fHarvard Medical School, Harvard University, Boston, MA, 02215 USA

ARTICLE INFO

Article history:

Received 29 March 2020

Accepted 7 December 2020

Available online 26 December 2020

2000 MSC: 41A05, 41A10, 65D05, 65D17

Keywords: congenital heart disease, multi-view learning, multi-channel networks, neural aggregation.

ABSTRACT

Congenital heart disease (CHD) is the most common birth defect and the leading cause of neonate death in China. Clinical diagnosis can be based on the selected 2D key-frames from five views. Limited by the availability of multi-view data, most methods have to rely on the insufficient single view analysis. This study proposes to automatically analyze the multi-view echocardiograms with a practical end-to-end framework. We collect the five-view echocardiograms video records of 1308 subjects (including normal controls, ventricular septal defect (VSD) patients and atrial septal defect (ASD) patients) with both disease labels and standard-view key-frame labels. Depthwise separable convolution-based multi-channel networks are adopted to largely reduce the network parameters. We also approach the imbalanced class problem by augmenting the positive training samples. Our 2D key-frame model can diagnose CHD or negative samples with an accuracy of 95.4%, and in negative, VSD or ASD classification with an accuracy of 92.3%. To further alleviate the work of key-frame selection in real-world implementation, we propose an adaptive soft attention scheme to directly explore the raw video data. Four kinds of neural aggregation methods are systematically investigated to fuse the information of an arbitrary number of frames in a video. Moreover, with a view detection module, the system can work without the view records. Our video-based model can diagnose with an accuracy of 93.9% (binary classification), and 92.1% (3-class classification) in a collected 2D video testing set, which does not need key-frame selection and view annotation in testing. The detailed ablation study and the interpretability analysis are provided.

The presented model has high diagnostic rates for VSD and ASD that can be potentially applied to the clinical practice in the future. The short-term automated machine learning process can partially replace and promote the long-term professional training of primary doctors, improving the primary diagnosis rate of CHD in China, and laying the foundation for early diagnosis and timely treatment of children with CHD.

1. Introduction

Congenital heart disease (CHD) is a cardiovascular malformation caused by abnormal development of the embryonic heart and vascular tissue (van der Linde *et al.*, 2011). It is the most common birth defect in China and the leading cause of neonate and child death (Dai *et al.*, 2011). The incidence of CHD is about 0.5-0.7% in China, among all subtypes, ventricular septal defect (VSD) and atrial septal defect (ASD) are the most common ones (Yang *et al.*, 2009; Wu *et al.*, 2014). Early screening and accurate diagnosis of CHD are essential to reduce the risk of this disease (Pruetz *et al.*, 2019), and most of the simple CHDs can be cured by timely surgery (WU *et al.*, 2006). With recent improvement of medical level in China, CHD has received increasing attention. Some pregnant women with environmental risk factors undergo fetal echocardiogram during pregnancy, which enables early diagnosis of this part of CHD (Zhang *et al.*, 2015; Zhao *et al.*, 2014).

Diagnosis of CHD during early childhood depends on multi-view echocardiograms (Sun *et al.*, 2015). Complex CHD often presents obvious echocardiogram abnormalities, which can be more easily recognized by doctors at primary hospitals, allowing a timely transfer of the patients to superior professional children's hospitals. However, owing to the lack of experienced cardiac sonographers, there remain a large number of children with delayed diagnosis of CHD, especially simple subtypes like ASD and VSD. The delayed or missed diagnosis may result in repeated pneumonia (Luo *et al.*, 2019), the missing of the best timing of surgery, making a serious impact on children's prognosis and future life (Chang *et al.*, 2008).

With sufficient data and the emergence of novel data driven learning algorithms, machine learning technology can make up for the missed diagnosis of CHD caused by the lack of professional doctors in echocardiogram (Litjens *et al.*, 2019). Recent researches on automated analysis of the abnormality of the heart structure usually focus on the single-view two-dimensional photographs or dynamic images of the echocardiogram (Zheng *et al.*, 2008; Pereira *et al.*, 2017; Maraci *et al.*, 2018). However, the clinical diagnosis of CHD obtained from a single view can be hardly reliable, therefore a multi-view joint diagnosis is necessary (Lai *et al.*, 2006).

In this study, we collect a dataset with 1308 children multi-view echocardiogram video records and the key-frame of each video is labeled by an experienced doctor. With the help of this well-organized dataset, we propose to make automated diagnosis in an end-to-end manner. Although the convolutional neural networks (CNN) have shown tremendous success in processing two dimensional (2D) echocardiograms, the processing of multi-view ultrasound images is under-explored (Liu *et al.*, 2019a).

A multi-channel CNN is constructed for the automated analysis of five views of ultrasound heart images, which summarizes the information in each view with the pointwise convolution. We explore the five views jointly to diagnose CHD. To the best of

our knowledge, this is the first effort to use five 2D echocardiograms views to assist with the deep learning-based automated diagnosis of CHD.

Moreover, it is well known that deep learning is data starved. The limited training data cannot support the training of network with a large number of parameters, and usually results in over-fitting. To alleviate this problem in our task, the Depthwise Separable Convolution (DSC) (Howard *et al.*, 2017) is adopted to largely reduce the network parameters. Other than the relatively limited training samples, the number of collected negative samples is usually larger than the number of positive (VSD and ASD) samples. We approach the imbalanced class problem by augmenting the positive training samples.

To alleviate the manual labeling of the key-frame, we further propose to process the raw videos without the need of key-frame labels and view class labels. We analyzed a series of aggregation framework to adaptively process the video data and achieve the comparable performance as the key-frame based version. Moreover, the key-frame labels in our dataset can also be utilized to guide the aggregation. Besides, a view detection module can be added to automatically classify the collected views and put them in order for the corresponding view-specific encoder.

This study aims to investigate the 2D echocardiograms analysis with five standard views. We collect the most common VSD and ASD as examples, and provide a serial of strong baseline methods. The contributions of this work can be summarized as follows:

- We collect the first large scale CHD dataset for five-view two-dimensional echocardiograms analysis. Both the raw video and the experienced doctor labeled key-frame versions are provided.
- A series of powerful baselines to explore both the key-frame-based and video-based multi-view diagnose. The depthwise separable convolution-based efficient multi-channel CNN architecture can utilize the limited data to achieve satisfactory multi-view diagnosis performance.
- In this paper, we propose a novel soft-attention framework to process video frames with or without the key-frame label in the training, while we do not need key-frame labeling and view labels in the testing stage. Four practical video aggregation schemes are investigated to explore the information among the variable number of frames.
- We also demonstrate that the proposed attention mechanism provides fine-scale attention maps that can be visualized, with minimal computational overhead, which helps with the interpretability of predictions.

2. Related work

Deep learning for echocardiograms. With the rapid development of machine learning, automated machine analysis and interpretation technology is increasingly used in medicine (Liu *et al.*, 2019a). More and more researchers are turning their attention to the data analysis and identification of medical images (De Fauw *et al.*, 2018; Esteva *et al.*, 2017; Bejnordi *et al.*, 2017).

*Corresponding to Binbin Wang (Tel: +86 10 62173443), Wanqing Xie (Tel: +1-617-230-8174), and Xin Zhang (Tel: +86 10 59616161)

¹Jing Wang and Xiaofeng Liu contributed equally to this work.

Table 1. The statistics of collected patient samples. VSD: ventricular septal defect; ASD: atrial septal defect. Testing set is consisted with 82 healthy control subjects, 21 ASD subjects and 28 VSD subjects.

	Healthy control	ASD	VSD
A4C	823	209	276
PSLAX of left ventricle	820	171	218
PSSAX of aorta	764	166	243
SXLAX of two atria	625	102	100
SSLAX of aortic arch	641	84	99

Early attempts extract the handcrafted features from a cardiovascular image (Maraci et al., 2017) and fed them to the statistical classifier, e.g., support vector machine (Cortes and Vapnik, 1995). The feature for CHD prediction is usually describing the texture and shape characteristics (Criminisi et al., 2009; Zheng et al., 2008). The kernel dynamic texture models have been applied to label each individual ultrasound image (Kwitt et al., 2013).

Recently, convolutional neural networks based deep learning gained enormous successes in image analysis tasks (Liu, 2020; Liu et al., 2019e). Instead of designing the feature by humans and subsequently feeding it to a prediction model, deep learning proposes to simultaneously learn relevant features and the prediction model from the raw image in an end-to-end fashion (LeCun et al., 2015; Che et al., 2019; He et al., 2020; Han et al., 2020; He et al., 2020).

Related studies have shown that using deep neural networks, machines can effectively identify abnormalities from ultrasonic views (Litjens et al., 2019; Liu et al., 2020a, 2019d, 2018b, 2020b). Diller et al. use a four-convolutional-layer CNN to process a single image to assess patients with a systemic right ventricle (Diller et al., 2019).

At present, researchers focus on two-dimensional photographs or dynamic images of the echocardiogram and explore the intelligent recognition method and diagnostic performance for abnormality of the heart structure (Zheng et al., 2008; Pereira et al., 2017; Maraci et al., 2018). The most closely related work is that of Zhang et al. (Zhang et al., 2018), which use two neural network branches to diagnose PSLAX, PSSAX, apical 2-chamber, apical 3-chamber, and A4C images for hypertrophic cardiomyopathy and cardiac amyloid.

Image set based recognition. A video can be considered as an image set with ordered images and has been actively studied. (Yang et al., 2017; Liu et al., 2017) compute a score for each image with neural image assessment modules. Then, a set of features are aggregated to a fixed size feature vector via weighted average pooling. Without inner-set interactions, this may result in redundancy and sacrifice the diversity in a set. (Liu et al., 2018a) proposes to exploit the inner-set relationship using reinforcement learning. However, its decision is based on the incomplete observation of the averaged features. Besides, the reinforcement learning itself is usually unstable (Henderson et al., 2017), the feature dimension is necessary to be compacted from 1024 in its GoogleNet backbone (Szegedy et al., 2016) to 128 which undoubtedly weakened its representation ability.

Self-attention and non-local. As attention models grew in

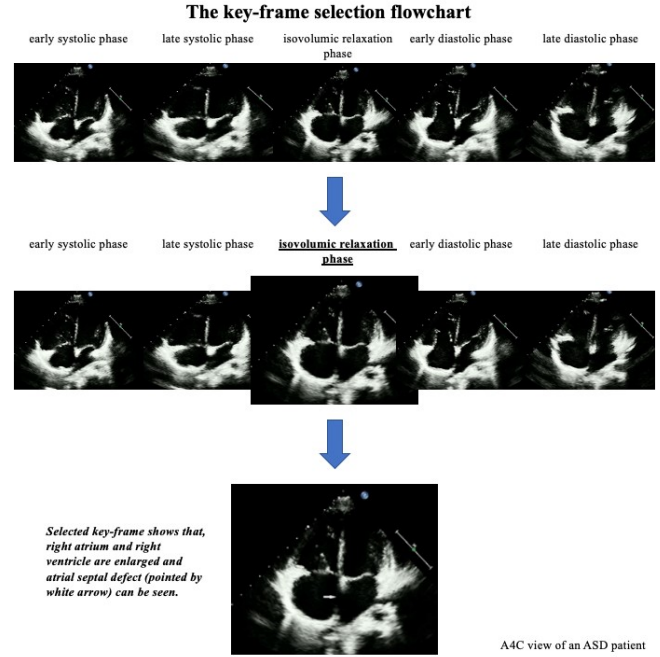


Fig. 1. The key frame selection flowchart. We take an A4C view of an ASD patient as an example.

popularity, (Vaswani et al., 2017) developed a self-attention mechanism for machine translation. It calculated the response at one position as a weighted sum of all positions in the sentences. A similar idea is also inherited in the non-local algorithm (Buades et al., 2005), which is a classical image denoising technique. The interaction networks also developed for modeling the pair-wise interactions (Battaglia et al., 2016; Hoshen, 2017; Watters et al., 2017; Yang et al., 2018). Moreover, (Wang et al., 2018) proposes to bridge self-attention to the more general class of non-local filtering operations. (Zhou et al., 2018) proposes to learn temporal dependencies between video frames at multiple time scales. Inspired by above works, we further adapted this idea to the multi-view echocardiograms video analysis with variable lengths.

Temporal modeling is widely used in video classification, video-based identification and temporal action detection etc. (Zhou et al., 2017) proposes to model the temporal information between frames using a recursive neural network (RNN). The 3D CNN is later developed to extract spatial-temporal features from video clips directly (Tran et al., 2015). However, it is not compatible with our attention-based module. Two types of temporal attention methods have been recently developed. The first method uses spatial convolution first followed by the fully connected (FC) layers (Liu et al., 2017). The FC layers limit the model to fixed length video clip. The second approach chooses the temporal convolution instead of FC (Gao and Nevatia, 2018). (Gao and Nevatia, 2018) uses temporal convolution for person re-identification with fixed length clip, we show that it is promising for echocardiograms video analysis with variable lengths.

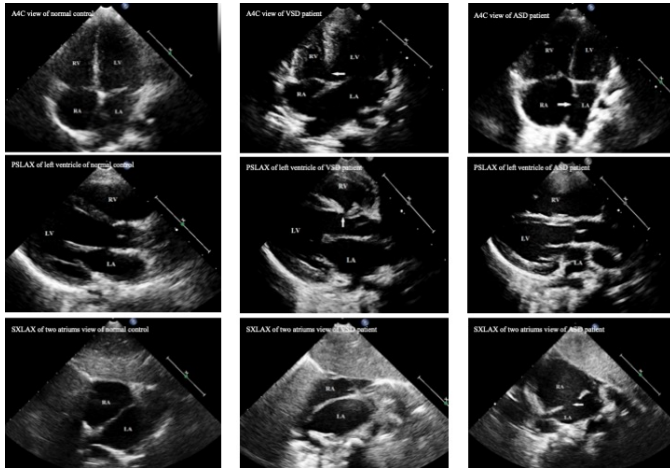


Fig. 2. The key frame selection results of different views. From left to right: A4C, PSLAX and SXLAX view. From top to bottom: normal controls, VSD patients and ASD patients.

3. Methodology

3.1. Standard collection of cardiac ultrasound image data

Totally 1308 children (823 healthy controls, 209 VSD and 276 ASD) are collected from Beijing Children's hospital. Each patient has the echocardiogram video records from 1 to 5 views that can be sufficient for diagnosis (shown in Table 4). Besides, a single high-quality 2D frame in a video of each view is selected by the doctor to construct a 2D echocardiogram dataset.

The patient was placed in the supine position and the chest was exposed for the echocardiogram. We used PHILIPS iE 33, iE Elite, and EPIQ 7C (Philips Electronics Nederland B.V.) as instruments. The transducers frequency was ranging from 3-8 MHz. According to the heart segmental approach, the heart position, atrial position, and ventricular position were determined, and the connection relationship between the atria, ventricle, and aorta was analyzed. The atrial septum and ventricular septum were observed for whether had defects, and cavity or pulmonary venous return was noted. Five standard 2D views, the parasternal long axis view (PSLAX) of left ventricle, the parasternal short-axis view (PSSAX) of aorta, the apical four chambers view (A4C), the subxiphoid long axis view (SXLAX) of two atria and the suprasternal long-axis view (SSLAX) of aortic arch were collected.

All patients' diagnosis was confirmed by either at least two senior ultrasound doctors or intraoperative final diagnosis. The study protocol was approved by the Ethics Committee of Beijing Children's Hospital (No. 2019-k-342).

The originally collected videos of each view are three cardiac cycles. To facilitate the processing, we also sampled a subset by randomly crop a clip of 0.8 seconds for each view. Since the cardiac cycle of the child is typically from 0.5-0.6s, the clip of 0.8 can usually have and only have one complete cardiac cycle and one key-frame. The originally collected video dataset has a different video length for different patients and different views.

Actually, the labeling of cardiac cycles is very costly in real-world implementations. Besides, taking many cardiac cycles as

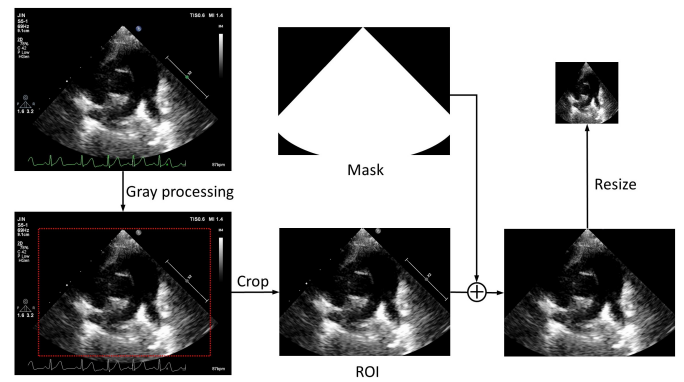


Fig. 3. Data pre-processing flowchart.

input will be very redundant. Considering we are using the same instrument for collection, the temporal resolution is consistent for all of the samples. The final dataset has two versions, i.e., the original video and the cropped video. We utilize the cropped version for our video-based experiments, while the original version is possible to support more sophisticated settings.

The key-frame is manually selected by the experienced doctor. We selected the isovolumic relaxation phase as a key-frame when the ventricles finish contracting and start to relax, the defects of VSD and ASD both could be shown clearly at that time. The flowchart in Fig. 1 could show the key-frame selection process, and the selected samples of different views are shown in Fig. 2.

We select 10% patients to construct our fixed testing set (i.e., 82 healthy control, 21 ASD and 28 VSD patients) in a patient-independent manner. That means the videos or keyframes used in training will not be incorporated in testing. Noticing that all our selected testing patients have all the five views original video records or selected key-frame.

3.2. Pre-processing

Each frame in the video or the selected keyframe is a two-dimensional echocardiogram with color instrument marks, which is essentially the three-channel RGB image. We first processed each frame to a single-channel grayscale image, and the region of interest (ROI) is cropped to remove the irrelevant parts. Because the collected ultrasonic image is a circular sector and some labels cannot be removed by rectangular cropping, a mask is designed to cover these factors. This is necessary, since only the positive samples have electrocardiogram record in our dataset and the CNNs can easily discriminate the samples according to this clue. To align with the input of CNNs, the masked ROI is resized to 128×128. The flow chart of pre-processing is shown in Figure 3. We use the same pre-processing for the image of each view and concatenate these five images following fixed order: PSLAX of left ventricle, PSSAX of aorta, A4C, SXLAX of two atria and SSLAX of aortic arch. Noticing that each view only has one image in our 2D echocardiogram dataset.

3.3. Key-frame based multi-view echocardiograms analysis

DSC for multi-view echocardiograms

Table 2. Multi-channel DSC network architecture. dw denotes the depth-wise convolution with size $H \times W$, and keep the same for N channels. pw denotes the point-wise convolution with size $1 \times 1 \times N$. We note that the first layer uses the conventional convolution with stride size 2 as in Howard et al. (2017).

Input Size	Type / Stride	Filter Shape
$128 \times 128 \times 5$	Conv / s2	32 kernels of $3 \times 3 \times 5$
$64 \times 64 \times 32$	Conv dw / s1	32 kernels of $3 \times 3 \times dw$
$64 \times 64 \times 32$	Conv pw / s1	64 kernels of $1 \times 1 \times 32 \text{ pw}$
$64 \times 64 \times 64$	Conv dw / s2	64 kernels of $3 \times 3 \times dw$
$32 \times 32 \times 64$	Conv pw / s1	128 kernels of $1 \times 1 \times 64 \text{ pw}$
$32 \times 32 \times 128$	Conv dw / s2	128 kernels of $3 \times 3 \times dw$
$16 \times 16 \times 128$	Conv pw / s1	128 kernels of $1 \times 1 \times 128 \text{ pw}$
$16 \times 16 \times 128$	Conv dw / s2	128 kernels of $3 \times 3 \times dw$
$8 \times 8 \times 128$	Conv pw / s1	128 kernels of $1 \times 1 \times 128 \text{ pw}$
$8 \times 8 \times 128$	Flatten	N/A
8192	FC1	1024
1024	FC2	128
128	Classifier	Softmax; 2 or 3-dim

It is well known that neural networks are trained to approximate a mapping function by concatenation of multiple-layer simple non-linear functions and that the deeper structure is usually more powerful with respect to representational ability. However, deep learning is data starved. Thus, the limited training data cannot support tuning of the network with a large number of parameters, and usually results in overfitting. To alleviate this problem in our congenital cardiovascular disease diagnosis task, the Depthwise Separable Convolution (DSC) (Howard et al., 2017) is adopted to largely reduce the network parameters.

We give a detailed comparison of the standard and depth-wise separable convolution in Figure 4. In DSC the convolution process is broken down into two operations: depth-wise convolutions and point-wise convolutions. In depth-wise operation, convolution is applied to a single channel at a time unlike standard CNN's in which it is done for all the N channels.

Noting that the DSC is originally designed for RGB data, which differs from our input, we inherit the idea of the DSC to construct our multi-channel CNN.

The widely used AlexNet (Simonyan and Zisserman, 2014) has 60 million to-be-learned parameters, while the DSC (with width multiplier 0.50) (Howard et al., 2017) can achieve comparable performance in the ImageNet object classification dataset using only 1.32 million parameters.

Network structure design

The diagnosis of congenital cardiovascular disease is more challenging than a common object recognition task because of its multi-view data structure as well as the relatively limited and unbalanced training sample.

The decision is based on five views, which can incorporate more complementary information than a single view-based diagnosis. However, this also introduces the challenge of information fusion. We propose to concatenate the five views sequentially and produce a matrix with size $128 \times 128 \times 5$. A five-channel CNN, as shown in Figure 5, is developed to take the concatenated matrix as input. Each DSC block incorporates a depthwise convolution and a pointwise convolution. After a few DSC layers, the feature maps will be flattened as the feature vector that will be processed by two fully connected layers with sizes 1024 and 128, respectively. The detailed network structure is given in

Table 3. Multi-branch DSC network architecture. Each channels are processed by the independent convolutional layers and concatenated in the first fully connected layer.

Input Size	Type / Stride	Filter Shape
Five $128 \times 128 \times 1$	Conv / s2	5-group of 32 kernels of $3 \times 3 \times 1$
Five $64 \times 64 \times 32$	Conv dw / s1	5-group of 32 kernels of $3 \times 3 \times dw$
Five $64 \times 64 \times 32$	Conv pw / s1	5-group of 64 kernels of $1 \times 1 \times 32 \text{ pw}$
Five $64 \times 64 \times 64$	Conv dw / s2	5-group of 64 kernels of $3 \times 3 \times dw$
Five $32 \times 32 \times 64$	Conv pw / s1	5-group of 128 kernels of $1 \times 1 \times 64 \text{ pw}$
Five $32 \times 32 \times 128$	Conv dw / s2	5-group of 128 kernels of $3 \times 3 \times dw$
Five $16 \times 16 \times 128$	Conv pw / s1	5-group of 128 kernels of $1 \times 1 \times 128 \text{ pw}$
Five $16 \times 16 \times 128$	Conv dw / s2	5-group of 128 kernels of $3 \times 3 \times dw$
Five $8 \times 8 \times 128$	Conv pw / s1	5-group of 128 kernels of $1 \times 1 \times 128 \text{ pw}$
Five $8 \times 8 \times 128$	Flatten	N/A
40960	FC1	5120
5120	FC2	128
128	Classifier	Softmax; 2 or 3-dim

Table 2.

We note that the multi-branch framework can be an alternative to the multi-channel network. In the multi-branch framework (Lee et al., 2016), each of the five views is passed through an independent neural network to create high-level features, which are then combined at the end of the network before making a final prediction.

Following the comparison setting in (Lee et al., 2016), we configure the convolutional layers for the view-independent branch, and the fully connected layers for the merged feature processing network. Considering the input sample of each branch has a size of $128 \times 128 \times 1$, we change the first convolutional layer to the 32 kernels with the size of $3 \times 3 \times 1$ kernels. With five branches, we have five $8 \times 8 \times 128$ to be fused feature maps. As in (Lee et al., 2016), we concatenate the five flattened 8192-dimensional vectors and followed by the first fully connected layer. The detailed network structure of multi-branch DSC is detailed in Table 3.

Our paper proposes to explore the CHD diagnosis with multi-view echocardiograms, and develop a powerful baseline model. The multi-channel scheme can adaptively learn the information fusion in each layer, instead of the simple concatenation in a fully connected layer (Lee et al., 2016). Actually, the different views of echocardiograms also share some similarities. Therefore, the filters learned in a view may potentially be helpful in the other view. Moreover, the multi-channel model with a single-branch can have significantly fewer parameters than the multi-branch counterpart. The fewer parameter is important for the limited training data. Besides, the memory cost of the multi-channel module is much fewer than the multi-branch module. We note that the multi-channel model is not applicable for heterogeneous inputs (e.g., image and EEG data), since they require different network structure for different inputs. With sufficiently training data, the performance of multi-branch and multi-channel are usually similar for homogeneous input (Lee et al., 2016).

For binary classification (negative or positive), we use the sigmoid function as the output unit and optimize the network with binary cross-entropy loss.

For the three-class classification (negative, ASD and VSD), we use the transfer learning from binary task. We choose the best pretrained binary classification model and configure its output layer as three neural units followed by a SoftMax function. The

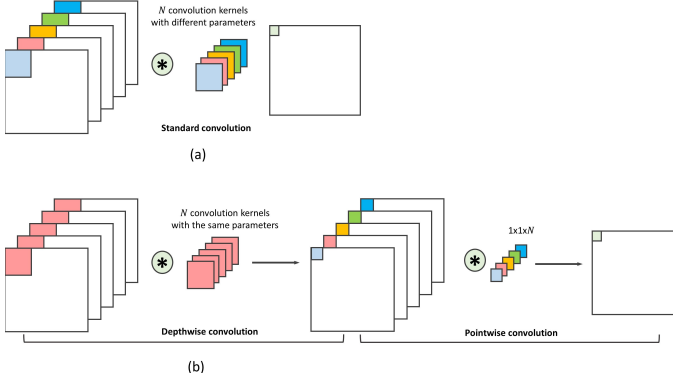


Fig. 4. Illustration of the convolution operation. a. the standard convolution; b. the Depthwise Separable Convolution.

network is fine-tuned with multiclass cross-entropy loss. To save time, the CNNs could initially be trained to solve the 3-class task and their outputs could be then merged to get the binary prediction results.

Balancing classes with data augmentation In the three-class classification case, the number of collected negative samples is almost four times larger than that of VSD or ASD. The richer source of the healthy control sample is common in many medical tasks. However, the unbalanced data distribution potentially results in the inference of network bias to the negative class. Therefore, we propose augmentation of the positive training samples by constructing “virtual” positive patients using a different combination of the original positive patients in the training set. We note that the testing set is always fixed with the original five views. Specifically, for the training samples without all five views, we randomly sample the missed views from the other patients in the same class (i.e., negative, ASD and VSD) to construct more patient samples. Since the original 1 to 4 views can be sufficient for diagnosis, the added view from the same class will not change the disease class. In our experimental settings, we construct 4 times positive (VSD+ASD) samples of the original positive patients to achieve a balanced training distribution.

3.4. Video based multi-view echocardiogram analysis

Keyframe selecting can always be costly, and the automatic analysis of multi-view videos in an end-to-end fashion is necessary when large scale samples are collected. Therefore, we propose to investigate the possible deep aggregation frameworks for our video-based multi-view two-dimensional echocardiogram analysis. The basic idea is to assign a larger weight for the more important frame in the diagnosis procedure.

The overall framework is shown in Figure 6. The video of each view is feed to a view-specific encoder to extract the frame-independent feature representations. For example, the k A4C frames x_{A4C}^k is encoded to k f_{A4C}^k feature vectors. This is a typical solution for dimension reduction and information compressing (Liu et al., 2019b, 2018a). Specifically, it embeds the images into latent space independently and generates the corresponding feature sets $\{f_{A4C}^k\}$, where $f_{A4C}^k \in \mathbb{R}^{H \times W \times D}$, $k \in 1, 2, \dots, K$ indicates the K frames, H , W and D are the height, width and

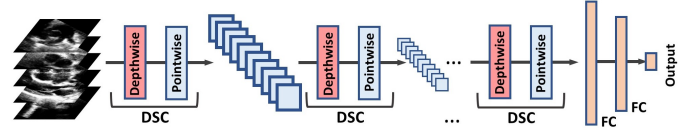


Fig. 5. Illustration of the multichannel convolutional neural networks for key frame-based multi-view diagnosis.

channel dimension of our representation, respectively. Empirically, we use the convolutional layers of key-frame based model DSC as our encoder (e.g., Enc_{A4C}), and the size of x_{A4C}^k is $H = 8$, $W = 8$, $D = 128$.

Then, a feature aggregation module is applied to the features of each frame and produces a representation of this view. The ordered representations for all five views are concatenated to the later disease classifier Cls_d for binary or three-class classification. We adapted four possible feature aggregation modules for our video-based echocardiogram analysis.

Frame-independent aggregation

The first is frame-independent aggregation (Yang et al., 2017), which uses a neural quality assessment module to assign a weight for each frame (Figure 9). Then, the features are averaged according to the weight. The f_{A4C}^k is flattened to an 8192-dimensional vector $\overline{f_{A4C}^k}$ and followed by the weight assignment network (Yang et al., 2017). We also chose the two-block neural network as in (Yang et al., 2017) for weight assignment.

The first attention block is the linear transformation with softmax normalization $a_{A4C}^k = \frac{\exp(e_{A4C}^k)}{\sum_j \exp(e_{A4C}^j)}$ and $e_{A4C}^k = q_{A4C}^T \overline{f_{A4C}^k}$, where $a_{A4C}^k \in \mathbb{R}$, vector q_{A4C} has the same size as a single feature $\overline{f_{A4C}^k}$. Then, the weighted sum of $\overline{f_{A4C}^k}$ with the weights a_{A4C}^k is calculated to form the aggregated feature $r_{A4C} = \frac{1}{K} \sum_k a_{A4C}^k \overline{f_{A4C}^k}$, where r_{A4C} is also a 8192-dimensional vector, k is the index of K frames.

With the linear aggregated r_{A4C} , the nonlinear block applies the tanh operation to calculate the final aggregation $\widetilde{f_{A4C}} = \tanh(w^T r_{A4C} + b)$, where w and b are the weight and bias term of linear neural unit.

The aggregated 8192-dimensional feature representation $\widetilde{f_{A4C}}$ is concatenated with the feature vector of the other views and followed by the fully connected layer. We also use the two-layer fully connected layers with the size of 1024 and 128 as in image-based classification settings. The frame-independent weighting scheme has a simple structure and each frame can be processed parallel. However, the assessment is only based on a single frame input and cannot consider the inter-frame relationship. Actually, the neighboring frames in the echocardiogram video are closely related and follows the cardiac cycle.

Recurrent neural network

The recurrent neural network (RNN) is a well-established method for video processing. Although its training can be unstable and the processing is relative slow (Yang et al., 2017; Liu et al., 2018a).

We propose to use RNN to assign the weight of each frame by considering the correlations with the neighboring frames. We adapt our video-based attention scheme based on RNN as shown

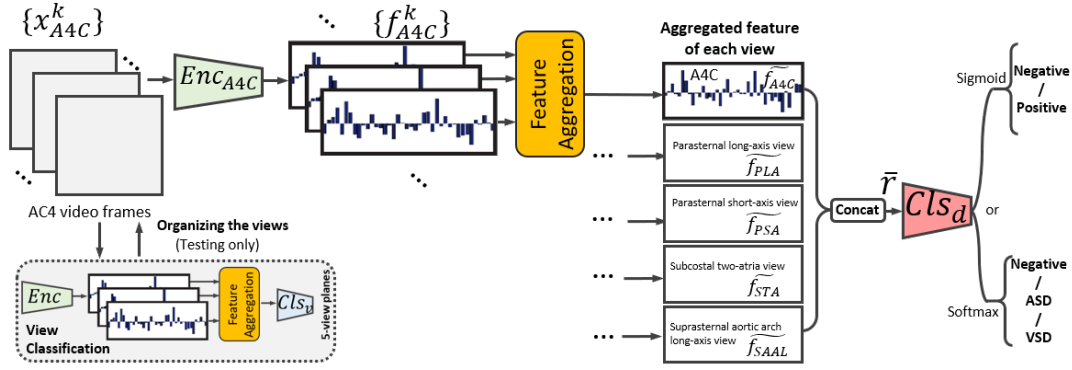


Fig. 6. The overall framework of video-based multi-view two-dimensional echocardiograms analysis. The view classification module adopts the same backbone as the view-independent encoder.

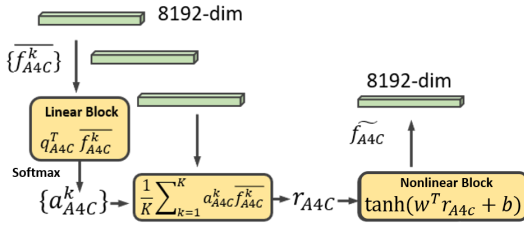


Fig. 7. The frame-independent aggregation module used for video-based echocardiogram analysis.

in Figure 8. The f_{A4C}^k is also flattened to an 8192-dimensional vector \bar{f}_{A4C}^k and followed by the RNN.

Specifically, the bi-directional long short-term memory network is employed as the recurrent layer, which takes the sequential feature vectors as inputs and produces a sequence of activations. Then, the activations are normalized via the softmax. It is a typical solution for sequential data, but usually takes much more processing time in both training and testing, and hard to train (Liu et al., 2019c). Actually, the cardiac cycle in echocardiogram videos is important for keyframe selection.

Non-local aggregation

Considering the training difficulty of RNN and information loss along with the recurrent scheme, we propose to utilize the non-local aggregation framework which can take an arbitrary number of frames as input and adaptively learn a complementary representation by considering the redundancy between each frame. As shown in Fig. 9, the non-local block receives K extracted features $\{f_{A4C}^k\}$ and restructuring them based on inner-set correlations. $\{f_{A4C}^k\}$ are deterministically computed from the image set $\{x_{A4C}^k\}$, they also inherit and display large variations and redundancy.

For K feature maps, there are $K \times 8 \times 8$ positions in the HW plane. Following (Wang et al., 2018), we use $i \in \{1, 2, \dots, K \times 8 \times 8\}$ to index the position whose response is to be computed, and j is the index that enumerates all of the other possible positions.

We omit the superscript k and subscript A4C for simply and use $f_i \in \mathbb{R}^{1 \times 1 \times 128}$ denote the feature vector at position i . After

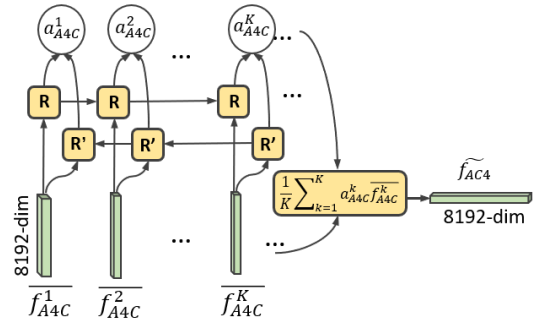


Fig. 8. The recurrent neural network module used for video-based echocardiogram analysis.

the to be processed f_i is chosen, there are $K \times 8 \times 8 - 1$ possible to be compared features which are denoted by $f_j \in \mathbb{R}^{1 \times 1 \times 128}$. The non-local block can be formulated as

$$f'_i = w \left[\frac{1}{C_i} \sum_{\forall j} e^{f_i^T f_j} g(x_i) \right] + f_i$$

$$C_i = \sum_{\forall j} e^{f_i^T f_j} \quad (1)$$

where the scalar $e^{f_i^T f_j} \in \mathbb{R}$ is the dot-product similarity of f_i and f_j . We can also use the Euclidean distance to model the relationship of f_i and f_j , but the dot-product similarity is more implementation-friendly in modern deep learning platforms. $g(f_i)$ is an embedding function, which applies the point-wise $1 \times 1 \times 128$ convolutions to f_j , and output a vector with the size of $1 \times 1 \times 128$. C_i is a normalization term. $w \in \mathbb{R}^{1 \times 1 \times 128}$ is a to be learned weight vector.

For a specific f_i , we traverse its $K \times 8 \times 8 - 1$ possible neighboring f_j to compute $\frac{1}{C_i} \sum_{\forall j} e^{f_i^T f_j} g(x_i)$. After the processing of a non-local block, f_i is reconstructed to $f'_i \in \mathbb{R}^{1 \times 1 \times 128}$. By traversing all $K \times 8 \times 8$ f_i , the non-local block output K feature maps, each with the size of $8 \times 8 \times 128$. Then, the global pooling (Wang et al., 2018) is applied for these K feature maps to calculate the element-wise average.

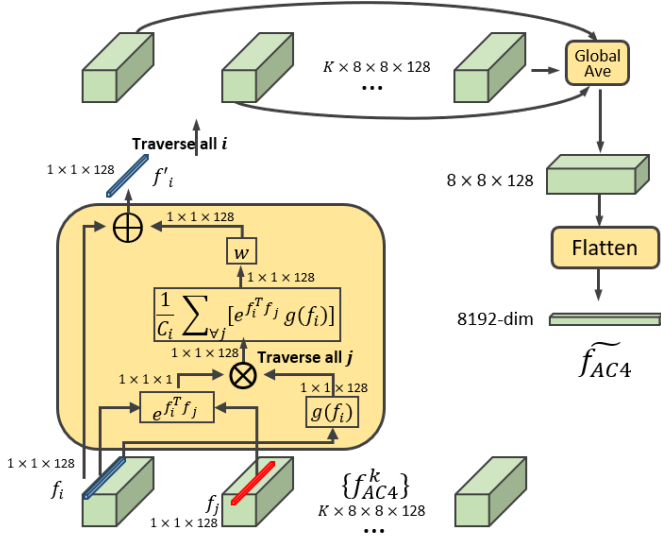


Fig. 9. The non-local aggregation module used for video-based echocardiogram analysis.

Temporal convolution Considering the complicated processing of the RNN and non-local modules, we propose to utilize the temporal convolution (Bai et al., 2018) to explore the neighboring relationships. Specifically, we adapt it to assign the attention score for aggregations as shown in Figure 10. It is introduced as a more efficient alternative for RNN (Gao and Nevatia, 2018).

Following (Gao and Nevatia, 2018), we first apply 64 spatial convolution kernels with the size of $8 \times 8 \times 128$ for each input feature map as the first layer to summarize the spatial information and produce a $1 \times 1 \times 128$ feature vector for each frame. We note that the spatial convolution is shared for all frames.

Then, we concatenate K feature vectors to a spatial-temporal feature map with the size of $k \times 1 \times 1 \times 128$. It is followed by two temporal convolution layers with padding and stride size is set to 1. The first layer uses 64 temporal convolution kernels with the size of $3 \times 1 \times 1 \times 128$, where 3 is for the time sequence direction, and the last three-dimension matches the size of the input frame-wise feature vector. Actually, each temporal convolution kernel output a K -dimensional vector which can be directly used as the attention score. However, the convolutional window for time direction is limited to three. To further enlarge the reception field in the time direction, we adopt the multiple temporal convolution layers as in (Bai et al., 2018). For the input of 64-dim features, i.e., the input matrix has the size $K \times 1 \times 1 \times 64$, the second layer temporal convolution kernel has the size of $3 \times 1 \times 1 \times 64$. The dimension of the output vector is equal to the number of frames in a video K . We use softmax to get the normalized attention score.

We note that the receptive field of temporal convolution is relatively small compared to the RNN. However, the redundancy in the video is mostly confined to neighboring frames. It can be a good balance of exploring the sequential relationship and processing speed.

Network and training details. The encoder and disease classifier Cls_d use the same backbone as the image-based setting (i.e.,

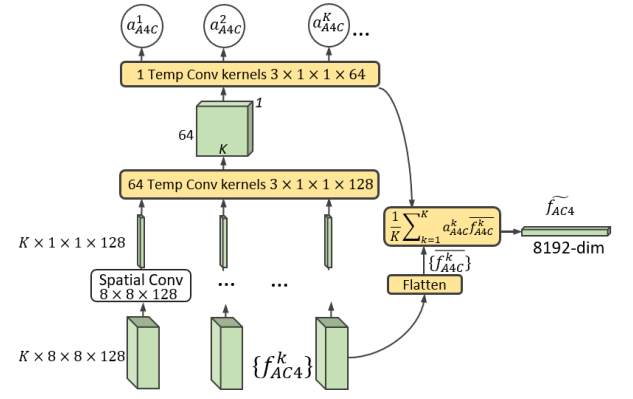


Fig. 10. The temporal convolution module used for video-based echocardiogram analysis.

AlexNet or DSC) for a fair comparison. The key-frame label in the training set can also be used as an additional signal to guide the neural aggregation modules, e.g., frame-independent aggregation, recurrent neural network, temporal convolution. Actually, the key-frame-based diagnosis is essentially a hard attention scheme that assigns 100% attention in the key-frame and 0% for the others. Since we know which frame should have a large weight (i.e., 100%), the difference between the predicted weights and the expert labeled weight can be used as an additional L2 loss for optimization.

3.5. View classification

Moreover, the view-specific encoder requires the collected five videos to be ordered with their view labels. A practical way is letting the user or clinician note the view in the collection stage or label the view based on collected echocardiograms. To alleviate the labeling task in real-world application and double-check the possible mislabeling of views, we propose an additional multi-view classification module.

The view classification module is used to predict the five echocardiograms views. Based on the view classification result, the five views are ordered and feed to their corresponding view-specific encoders.

The backbone structure of view classification module is the same as the normal/patient classifier, but without the multi-view feature concatenation and the Cls_v has five output units for five views instead of 2 or 3. We note that the aggregation in view-classification is the same as the diagnosis network in all of our network settings.

We note that the dataset has one video for each view. In some cases, more than one video may be classified to the same view and some of the other views are empty. For example, two videos' Cls_v softmax prediction has the maximum probability for the PSLAX view. Considering the softmax probability well calibrates the prediction confidence Zou et al. (2019), we compare the maximum probability and choose the larger one as the PSLAX view, while another video is assigned to the empty view.

The view classification for image-based diagnose is relatively redundant, since the user/doctor needs to know the view, and then make the key-frame selection. With the view classifier, our

Table 4. Performance of the disease classification task. The threshold for binary classification is 0.5, while we use *argmax* function to choose the largest probability in softmax output as the prediction of 3-class classification. VSD: ventricular septal defect; ASD: atrial septal defect; A4C: the apical four chambers view; DSC: Depthwise Separable Convolution.

Methods	Binary classification (Negative / VSD+ASD)		3-class classification (Negative / VSD / ASD)
	ACC	AUC	ACC
AlexNet (A4C view only)	0.840	0.806	0.817
AlexNet (Multi-branch)	0.893	0.843	0.876
AlexNet	0.895	0.845	0.878
DSC (A4C view only)	0.885	0.813	0.855
DSC (Multi-branch)	0.946	0.918	0.916
DSC ($\frac{1}{5}$ Multi-branch)	0.935	0.904	0.907
DSC	0.947	0.918	0.917
DSC (w/o transfer learning)	0.939	0.924	0.908
DSC+Augmentation	0.954	0.942	0.923

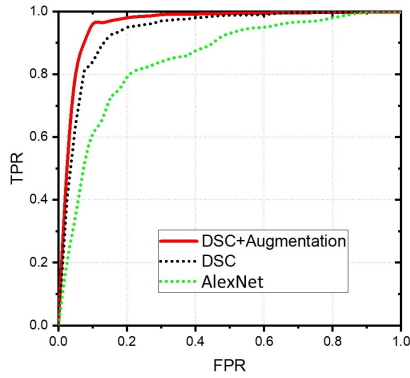


Fig. 11. Comparison of the DSC and AlexNet baseline with respect to ROC curves in the binary classification setting, when using five convolutional layers and a 128×128 input image size.

video-based diagnose system can take unordered five videos as input and automatically make the prediction.

4. Experiments and results

4.1. 2D key-frame testing set analysis

To investigate the diagnostic performance, we evaluate the neural network system in both binary and three-class classification settings. We choose the AlexNet as our baseline, which is a widely used structure but has significantly more parameters. As shown in Table 4, use of the DSC as the backbone outperforms the AlexNet baseline by 5.4% with respect to testing accuracy in the binary classification setting. More appealing, DSC is 45 times smaller and 9.4 times less compute than AlexNet. The positive data augmentation can efficiently improve the performance without sophisticated algorithms.

We note that the results of multi-channel network is comparable or even better than the multi-branch model. Moreover, the multi-channel model has a fifth of parameters in the convolutional layers.

Moreover, we propose to configure a multi-branch DSC network that has roughly the same parameters of our multi-channel DSC. Specifically, we use 7, 7, 14, 14, 28, 28, 28, 28, 28 kernels for each DSC layers instead of 32, 32, 64, 64, 128, 128,

Table 5. Confusion matrix of the key-frame-based binary classification task with DSC+Augmentation when setting the threshold with respect to $\bar{y}=0.5$.

	y=Negative	y=Positive
\bar{y} = Positive	0.061	0.980
\bar{y} = Negative	0.939	0.020

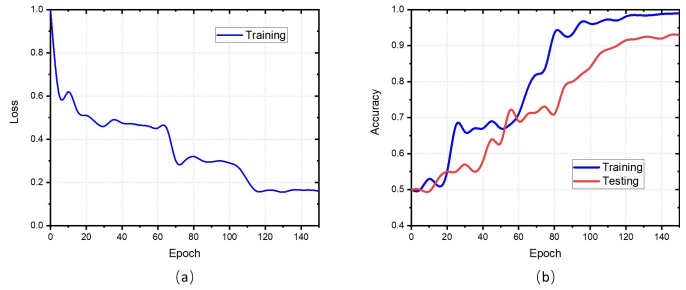


Fig. 12. The accuracy with respect to the number of epochs for our DSC. a. training loss; b. training validation accuracies.

128, 128, 128 respectively. Moreover, we set the first fully connected layer to the 1024-dimension. We denote this setting as DSC($\frac{1}{5}$ Multi-branch) in Table 4, which is significantly worse than the multi-channel DSC. Largely reduce the parameters can significantly affect the expressibility of the neural network and result in the performance drop of the vanilla multi-branch network.

By pre-train the network with binary classification and then modifying the sigmoid output unit to a three-way softmax unit, we can re-train the network in the three-class classification setting. The accuracy of three-class classification is usually lower than the binary counterpart since more fine-grained discrimination is required. While the pre-training with binary classification can be helpful for the performance.

To evidence the effectiveness of using five views as input rather than only using A4C view, we also propose the performance of a baseline that only use A4C view image as a single-channel input to AlexNet and our DSC networks. The improvements of the accuracy are more than 6.2%. We note that the proposed data augmentation does not apply to the single-channel case. In Table 4, we show the present model can diagnose positive or negative sample with an accuracy of 95.4%, 3-class classification (negative, ASD and VSD) with an accuracy of 92.3%. From the confusion matrix in Table 5, we can see that 98.0% of ASD/VSD samples are correctly diagnosed as positive.

The receiver-operating characteristic (ROC) curve plots the true positive rate (TPR) against the false positive rate (FPR) by varying a threshold. It illustrates the diagnostic ability of a binary classification system. A larger area under the curve indicates better performance. From Figure 11, we can see that the DSC outperforms the AlexNet baseline by a large margin, and that the data augmentation can further improve the performance consistently.

The loss value and training validation accuracies concerning the number of epochs are shown in Figure 12. The DSC converges to a steady optimum after about 120 epochs. Moreover,

Table 6. The video-based binary and 3-class classification accuracy and testing time. †Using the ground truth view label directly. *The key-frame annotation in the training set is used for the aggregation module training.

	Accuracy (Binary)	AUC (Binary)	Accuracy (3-class)	Testing time (Titan X)
Frame-independent	0.908	0.894	0.893	98ms
Frame-independent†	0.913	0.898	0.901	42ms
Frame-independent*	0.915	0.902	0.903	98ms
RNN	0.918	0.909	0.912	1537ms
RNN†	0.921	0.912	0.916	716ms
RNN*	0.931	0.917	0.918	1537ms
Non-local	0.933	0.918	0.914	294ms
Non-local†	0.935	0.920	0.918	187ms
Temporal convolution	0.939	0.922	0.921	132ms
Temporal convolution†	0.944	0.926	0.923	60ms
Temporal convolution*	0.947	0.928	0.925	132ms

Table 7. Confusion matrix of the video-based binary classification task with DSC+Augmentation (without key-frame label for training) when setting the threshold to $\bar{y} = 0.5$.

Frame-independent aggregation	y = Negative	y = Positive
\bar{y} =Positive	0.073	0.918
\bar{y} =Negative	0.902	0.082
Recurrent neural network	y = Negative	y =Positive
\bar{y} =Positive	0.085	0.939
\bar{y} =Negative	0.914	0.061
Non-local aggregation	y = Negative	y =Positive
\bar{y} =Positive	0.080	0.939
\bar{y} =Negative	0.920	0.061
Temporal convolution	y = Negative	y =Positive
\bar{y} =Positive	0.073	0.959
\bar{y} =Negative	0.927	0.041

the training becomes saturated in the middle stage, but does not settle down at the saddle point, which can be attributed to the Adam optimizer. The DSC yields reliable validation accuracies without overfitting to the training data.

4.2. The video-based model analysis

The video-based classification accuracy and testing time are compared in Table 6. The frame-independent aggregation is the fastest method, while the temporal convolution achieves the best accuracy and still has fast processing speed (132ms). We note that without the view classification module and use the ground truth view label, the processing can be 60ms. We also give the confusion matrix of four aggregation methods in the Table 7.

RNN can outperform the frame-independent aggregation, since it can take the other frames into account. However, its sequential processing can be much slower than the compared methods. The non-local scheme can utilize the powerful parallel computing ability of GPU and achieve comparable or even better performance than RNN with about five times faster processing. The non-local scheme considers all of the positions equally, while the neighboring frames in the echocardiogram video are closely related and follow the cardiac cycle. By considering this property, the temporal convolution can be a good solution and

Table 8. Five-view classification accuracy with different aggregation module. The correct classification of a subject indicates all five views are correctly classified.

Aggregation	Frame-independent	Non-local	RNN	Temp
Accuracy	0.992	0.995	0.991	0.994

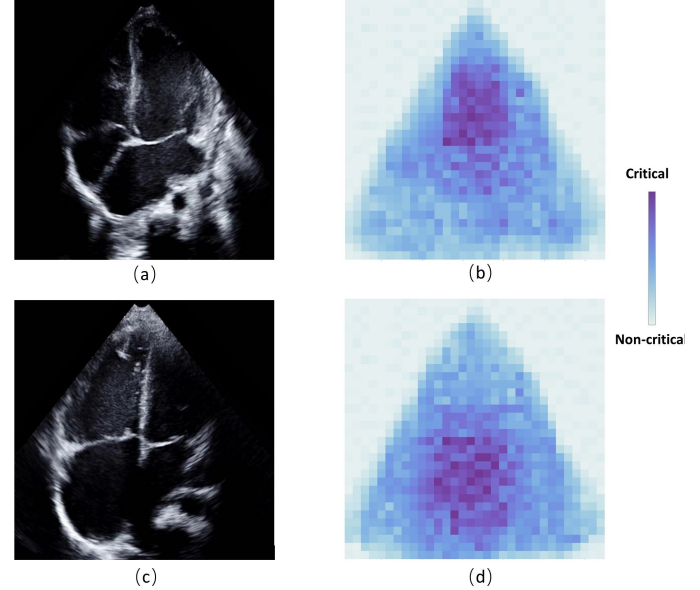


Fig. 13. The relative importance of different parts associated with the diagnosis. a. the apical four-chamber view of ASD patient; b. its corresponding relative importance visualization of different parts associated with ASD; c. the apical four-chamber view of VSD patient; d. its corresponding relative importance visualization of different parts associated with VSD.

achieves a good balance between the accuracy and processing speed.

Moreover, the frame-independent aggregation, RNN, and temporal convolution module can utilize the key-frame label. With the expert labeled keyframe, the performance of the aggregation module can be further improved.

Using the view classification module can achieve comparable performance as using the ground-truth view label. In Table 8, the view detection performance is compared. All of the aggregation methods can well maintain the view information and potentially facilitating the data collection in real-world implementations.

4.3. Understanding the perception area

Additionally, to understand the behavior of our neural network and identify regions for concatenated input that are critical for medical diagnosis, we propose an image occlusion analysis on our DSC model. We slide a box of $4 \times 4 \times 1$ zero-valued pixels along with the layer of the A4C in concatenated ultrasonic images from a patient that was correctly labeled as VSD or ASD samples by our trained model. As a consequence, the importance of each area can be characterized by the relative confidence of the samples being classified as positive. The resulting heat map is shown in Figure 13. The intensity of the heat map indicates the relative importance of each pixel block. The dark areas decrease the confidence of the model, suggesting that they are critical

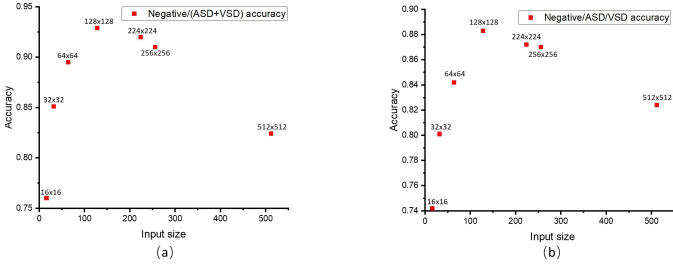


Fig. 14. Classification accuracy with different input image sizes. a. binary classification; b. three-class classification.

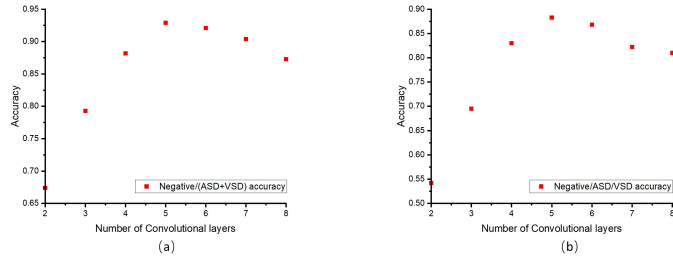


Fig. 15. Classification accuracy with different convolutional layers. a. binary classification; b. three-class classification.

areas for the diagnosis of cardiopathy. The dark blue regions in Figure 13 coincide with the ROIs of the VSD and ASD in the ultrasonic images, respectively.

4.4. Ablation study

The input size can significantly affect system performance. We further analyze the validation accuracy when choosing different input sizes. A low-resolution sample with too-small size (e.g., 16×16) cannot offer sufficient information. Meanwhile, a large size sample may require more neural network parameters and training data. The size 128×128×5 consistently achieves the best performance in our binary and three-class classification settings as shown in Figure 14.

Using more convolution layers can improve the representation ability of neural networks, but the additional parameters require significantly more training data. As shown in Figure 15, when we evaluate the performance with two to eight convolutional layers, respectively, the five-layer setting can well balance the number of parameters and training samples in our dataset.

5. Discussion

Clinical diagnosis of CHD is usually based on the comprehensive analysis of multiple views, and the disease diagnosis is a challenging task for the primary doctors. This study intends to develop a practical model to efficiently fuse the information in different views with either image or video modality. Based on clinical experience, we collect the five-view echocardiogram videos from 1308 children, with fine-grained key-frame labels.

As we have seen in the results, the 2D key-frame model can diagnose CHD or negative samples with an accuracy of 0.954, and in negative, VSD or ASD classification with an accuracy

of 0.923. Moreover, the video-based model can diagnose with an accuracy of 0.939 (binary classification), and 0.921 (3-class classification) in a collected 2D video testing set, which does not need key-frame selection and view annotation in testing. The present model developed has high diagnostic rates for VSD and ASD that can be potentially applied to the clinical practice in the future.

5.1. Clinical insights of multi-view analysis for CHD

The multi-view diagnosis accuracy can be 0.05 higher than the diagnose with only the A4C view as shown in Table 4. The AUC metric also coincides with the accuracy. To the best of our knowledge, this is the first effort to explore the five-view echocardiogram, and the corresponding dataset.

Previous researches on the use of artificial intelligence-assisted image processing mainly involved single view research, which can show the application value of artificial intelligence to a certain extent (Zheng et al., 2008). However, the clinical diagnosis of CHD cannot be reliably obtained from a single view. Instead, a multi-view joint diagnosis is necessary. A complete examination requires that the cardiovascular structures be imaged from multiple orthogonal planes. Therefore, we chose five standard 2D views, PSLAX of left ventricle, PSSAX of aorta, A4C, SXLAX of two atria and SSLAX of aortic arch, which in combination could describe all of the major cardiovascular structures in sequence (Lai et al., 2006). This study is the first to use five 2D views to assist with the diagnosis of CHD. Fortunately, our study showed that artificial intelligence and deep learning of the five standard 2D views could make identification and classification diagnosis of CHD.

The ultrasound manifestations of CHD can be divided into direct signs and indirect signs. Taking ASD as an example, the direct signs are that the echo is interrupted in the A4C and the SXLAX of two atria, and the blood flow communicates between the two atria by the defect. The indirect signs are that the right ventricle is enlarged in the PSLAX of left ventricle, the right atrium and ventricle are enlarged in the A4C, and the main pulmonary artery is widened in the PSSAX of aorta. For VSD, the direct signs are that the echo is interrupted in the A4C and the PSLAX of left ventricle, and the blood flow communicates between the two ventricles through the defect. The typical VSD indirect signs are that left atrium and ventricle are enlarged in the PSLAX of left ventricle and the A4C. However, when VSD is combined with pulmonary hypertension, it can also manifest as right atrium and ventricle enlargement and pulmonary artery widening which can be easily confused with ASD. At this time, PSLAX of left ventricle and the A4C show the ventricular defect, and the SXLAX of two atria show the completed atrial septum. These performances can help us to differentiate between the diagnosis of VSD and ASD, thereby showing the importance of multi-view joint diagnosis.

Furthermore, the appearance of CHD is various. For example, the defect site of VSD is variable, and defects at different positions will be displayed on different views. The defect can be shown in the A4C when it is located in the muscle and perimembrane, shown in the PSLAX of left ventricle when it is located in the infracristal, and shown in the PSSAX of aorta when it

is located in the subarterial, respectively. So comprehensive judgment of multi-view images is required to achieve the correct diagnosis.

5.2. Key-frame based multi-view analysis

In this work, we have investigated both the multi-channel and multi-branch CNN models with AlexNet and DSC backbones to aggregate the information in different views.

The single branch multi-channel structure can efficiently incorporate all the five views and correlate each channel via the pointwise convolution. Compared with the multi-branch model which fuse the information with the simple concatenation, the multi-channel model is possible to adaptively fuse the information in all of the layers.

Targeting for the limited training data, the DSC is adopted which compresses the parameter used in our AlexNet baseline by more than 32 times. To alleviate the class imbalance problem in the diagnosis task, the virtual patients are constructed by randomly sampling the missed view from the different patients as data augmentation. Its lite structure makes it promise to be deployed on many micro-embedded systems at a low cost. With the DSC, data augmentation and transfer learning for either binary classification or three-class classification, our method requires less training data and found excellent performance.

As investigated in our ablation study (Fig. 14), the spatial size of 128×128 can be an ideal choice of the input size. The cardiopathy region is usually a hole that requires sufficiently spatial resolution for diagnosing. While the larger input is found to be redundant. The choice of five convolution layer is based on our dataset. With more training data, we may expect a more representative model using deeper networks.

5.3. Video-based multi-view analysis

The collected raw ultrasonic data is video, i.e., a sequence of images. The key-frame selection is required for the 2D standard view-based solution, which can be tedious in implementation and hard for primary doctors. Actually, the video should cover the information of the key-frame, since the key-frame is a sub-set of the video. To alleviate the manually labeling task, we proposed a soft-attention framework to aggregate the information from these sequential frames.

We investigated four possible aggregation methods to adaptively assign the attention in each frame from the diagnosis perspective rather than the general image quality (e.g., blurry level). We systematically analyzed their performance and processing speed in Table 6.

The frame-independent aggregation is the fastest method, while the temporal convolution achieves the best accuracy and still has fast processing speed (132ms). RNN can take the other frames into account, but its sequential processing can be much slower than the other solutions. With parallel computing the non-local scheme can speed up the processing, but it redundantly considers all of the positions equally. However, the echocardiogram video follows the cardiac cycle, in which the frame is closely related to its neighboring frames. The temporal convolution explores the neighboring frames with its proper reception

fields and the parallel convolution operation, which achieves the real-time processing without sacrificing the accuracy.

With the key-frame annotation in the training samples, the aggregation framework, especially the temporal convolution scheme, can achieve comparable accuracy than the 2D standard view-based solution.

The five views should follow the fixed order of views to constitute an input of our model. As shown in Table 8, our multi-task model can detect the view with high accuracy and order them for the diagnosis. Actually, each view has significant patterns that can usually easy to detect. The performance gap of using view classifier and ground-truth view label is marginal in Table 7. The doctors in primary hospitals can collect the videos in random order without the view annotation. Moreover, the view classifier can also be used to double-check the clinical records.

5.4. Clinical prospects

The model developed in this study has high diagnostic rates for VSD and ASD and can be potentially applied to the clinical practice in the future, especially in primary hospitals. For example, 2D standard views are selected by a doctor, and a preliminary diagnosis of the echocardiogram image can be performed by an end-to-end multi-view deep learning system. Then the positive cases can be referred to specialist hospitals for further treatment. The short-term automated machine learning process can partially replace and promote the long-term professional training of primary doctors, improving the primary diagnosis rate of CHD in China, and laying the foundation for early diagnosis and timely treatment of children with CHD.

5.5. Limitations and future directions

Our model requires the input of five videos from five pre-defined views. We plan to construct a free-hand diagnosis system by detecting the view of each frame and aggregate the frames from the same view automatically. Moreover, a feature work can be using active learning to instruct the scanning which may reduce the number of views to be collected in real-world implementations. Thirdly, we will gradually include more echocardiograms of patients with other CHD subtypes in the following study, so that this diagnostic system can complete the diagnosis of more CHD subtypes in the future.

6. Conclusion

This study proposes to automatically analyze the five-view echocardiograms with end-to-end neural networks. Based on the collected multi-view dataset with both disease labels and standard-view key frame labels, our model can make the diagnosis on either selected 2D standard views or original videos. The model has high diagnostic rates for VSD and ASD and can be potentially applied to the clinical practice in the future, especially in primary hospitals. For example, 2D standard views are selected by a doctor, and a preliminary diagnosis of the echocardiogram image can be performed by artificial intelligence. Then the positive cases can be referred to specialist hospitals for further treatment. The short-term automated machine learning

process can partially replace and promote the long-term professional training of primary doctors, improving the primary diagnosis rate of CHD in China, and laying the foundation for early diagnosis and timely treatment of children with CHD.

Acknowledgements

This work was supported by the National Natural Science Foundation of China [grant number 91846102], the Fundamental Research Funds for the Central Universities [grant number GK2240260006], Beijing Municipal Administration of Hospitals Youth Programme [grant number QML20191208] and Jiangsu Youth Programme [grant number BK20200238].

References

- Bai, S., Kolter, J.Z., Koltun, V., 2018. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*.
- Battaglia, P., Pascanu, R., Lai, M., Rezende, D.J., et al., 2016. Interaction networks for learning about objects, relations and physics, in: *Advances in neural information processing systems*, pp. 4502–4510.
- Bejnordi, B.E., Veta, M., Van Diest, P.J., Van Ginneken, B., Karssemeijer, N., Litjens, G., Van Der Laak, J.A., Hermesen, M., Manson, Q.F., Balkenhol, M., et al., 2017. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama* 318, 2199–2210.
- Buades, A., Coll, B., Morel, J.M., 2005. A non-local algorithm for image denoising, in: *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, IEEE. pp. 60–65.
- Chang, R.K.R., Gurvitz, M., Rodriguez, S., 2008. Missed diagnosis of critical congenital heart disease. *Archives of pediatrics & adolescent medicine* 162, 969–974.
- Che, T., Liu, X., Li, S., Ge, Y., Zhang, R., Xiong, C., Bengio, Y., 2019. Deep verifier networks: Verification of deep discriminative models with deep generative models. *arXiv preprint arXiv:1911.07421*.
- Cortes, C., Vapnik, V., 1995. Support-vector networks. *Machine learning* 20, 273–297.
- Criminisi, A., Shotton, J., Bucciere, S., 2009. Decision forests with long-range spatial context for organ localization in ct volumes, in: *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 69–80.
- Dai, L., Zhu, J., Liang, J., Wang, Y.P., Wang, H., Mao, M., 2011. Birth defects surveillance in china. *World journal of pediatrics* 7, 302.
- De Fauw, J., Ledsam, J.R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Glorot, X., O'Donoghue, B., Visentin, D., et al., 2018. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine* 24, 1342–1350.
- Diller, G.P., Babu-Narayan, S., Li, W., Radojevic, J., Kempny, A., Uebing, A., Dimopoulos, K., Baumgartner, H., Gatzoulis, M.A., Orwat, S., 2019. Utility of machine learning algorithms in assessing patients with a systemic right ventricle. *European Heart Journal-Cardiovascular Imaging* 20, 925–931.
- Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S., 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542, 115–118.
- Gao, J., Nevatia, R., 2018. Revisiting temporal modeling for video-based person reid. *arXiv preprint arXiv:1805.02104*.
- Han, Y., Liu, X., Sheng, Z., Ren, Y., Han, X., You, J., Liu, R., Luo, Z., 2020. Wasserstein loss-based deep object detection, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.
- He, G., Liu, X., Fan, F., You, J., 2020. Image2audio: Facilitating semi-supervised audio emotion recognition with facial expression image, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.
- Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., Meger, D., 2017. Deep reinforcement learning that matters. *arXiv preprint arXiv:1709.06560*.
- Hoshen, Y., 2017. Vain: Attentional multi-agent predictive modeling, in: *Advances in Neural Information Processing Systems*, pp. 2701–2711.
- Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H., 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.
- Kwitt, R., Vasconcelos, N., Razzaque, S., Aylward, S., 2013. Localizing target structures in ultrasound video—a phantom study. *Medical image analysis* 17, 712–722.
- Lai, W.W., Geva, T., Shirali, G.S., Frommelt, P.C., Humes, R.A., Brook, M.M., Pignatelli, R.H., Rychik, J., 2006. Guidelines and standards for performance of a pediatric echocardiogram: a report from the task force of the pediatric council of the american society of echocardiography. *Journal of the American Society of Echocardiography* 19, 1413–1430.
- LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. *nature* 521, 436–444.
- Lee, D., Lee, J., Kim, K.E., 2016. Multi-view automatic lip-reading using neural network, in: *Asian conference on computer vision*, Springer. pp. 290–302.
- van der Linde, D., Konings, E.E., Slager, M.A., Witsenburg, M., Helbing, W.A., Takkenberg, J.J., Roos-Hesselink, J.W., 2011. Birth prevalence of congenital heart disease worldwide: a systematic review and meta-analysis. *Journal of the American College of Cardiology* 58, 2241–2247.
- Litjens, G., Ciompi, F., Wolterink, J.M., de Vos, B.D., Leiner, T., Teuwen, J., Išgum, I., 2019. State-of-the-art deep learning in cardiovascular image analysis. *JACC: Cardiovascular Imaging* 12, 1549–1565.
- Liu, S., Wang, Y., Yang, X., Lei, B., Liu, L., Li, S.X., Ni, D., Wang, T., 2019a. Deep learning in medical ultrasound analysis: a review. *Engineering*.
- Liu, X., 2020. Disentanglement for discriminative visual recognition. *arXiv preprint arXiv:2006.07810*.
- Liu, X., B.V.K. K., Yang, C., Tang, Q., You, J., 2018a. Dependency-aware attention control for unconstrained face recognition with image sets, in: *European Conference on Computer Vision*.
- Liu, X., Fan, F., Kong, L., Xie, W., Lu, J., You, J., 2020a. Unimodal regularized neuron stick-breaking for ordinal classification. *Neurocomputing*.
- Liu, X., Guo, Z., Jia, J., Kumar, B., 2019b. Dependency-aware attention control for imageset-based face recognition, in: *ArXiv*.
- Liu, X., Guo, Z., Li, S., You, J., B.V.K. K., 2019c. Dependency-aware attention control for unconstrained face recognition with image sets, in: *ICCV*.
- Liu, X., Han, X., Qiao, Y., Ge, Y., Li, S., Lu, J., 2019d. Unimodal-uniform constrained wasserstein training for medical diagnosis, in: *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 0–0.
- Liu, X., Xing, F., Yang, C., Kuo, C.J., El Fakhri, G., Woo, J., 2020b. Symmetric-constrained irregular structure inpainting for brain mri registration with tumor pathology, in: *MICCAI BrainLes*.
- Liu, X., Zou, Y., Che, T., Ding, P., Jia, P., You, J., Kumar, B.V., 2019e. Conservative wasserstein training for pose estimation, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.
- Liu, X., Zou, Y., Song, Y., Yang, C., You, J., K Vijaya Kumar, B., 2018b. Ordinal regression with neuron stick-breaking for medical diagnosis, in: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 0–0.
- Liu, Y., Yan, J., Ouyang, W., 2017. Quality aware network for set to set recognition, in: *Proc. IEEE Int. Conf. Comput. Vision Pattern Recognit.*, pp. 5790–5799.
- Luo, H., Qin, G., Wang, L., Ye, Z., Pan, Y., Huang, L., Luo, W., Guo, Q., Peng, Y., Wang, E., 2019. Outcomes of infant cardiac surgery for congenital heart disease concomitant with persistent pneumonia: A retrospective cohort study. *Journal of cardiothoracic and vascular anesthesia* 33, 428–432.
- Maraci, M.A., Bridge, C.P., Napolitano, R., Papageorgiou, A., Noble, J.A., 2017. A framework for analysis of linear ultrasound videos to detect fetal presentation and heartbeat. *Medical image analysis* 37, 22–36.
- Maraci, M.A., Xie, W., Noble, J.A., 2018. Can dilated convolutions capture ultrasound video dynamics?, in: *International Workshop on Machine Learning in Medical Imaging*, Springer. pp. 116–124.
- Pereira, F., Bueno, A., Rodriguez, A., Perrin, D., Marx, G., Cardinale, M., Salgo, I., del Nido, P., 2017. Automated detection of coarctation of aorta in neonates from two-dimensional echocardiograms. *Journal of Medical Imaging* 4, 014502.
- Pruetz, J.D., Wang, S.S., Noori, S., 2019. Delivery room emergencies in critical congenital heart diseases, in: *Seminars in Fetal and Neonatal Medicine*, Elsevier. p. 101034.
- Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- Sun, R., Liu, M., Lu, L., Zheng, Y., Zhang, P., 2015. Congenital heart disease: causes, diagnosis, symptoms, and treatments. *Cell biochemistry and biophysics* 72, 857–860.
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking

- the inception architecture for computer vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826.
- Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M., 2015. Learning spatiotemporal features with 3d convolutional networks, in: Proceedings of the IEEE international conference on computer vision, pp. 4489–4497.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need, in: Advances in Neural Information Processing Systems, pp. 5998–6008.
- Wang, X., Girshick, R., Gupta, A., He, K., 2018. Non-local neural networks, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
- Watters, N., Zoran, D., Weber, T., Battaglia, P., Pascanu, R., Tacchetti, A., 2017. Visual interaction networks: Learning a physics simulator from video, in: Advances in Neural Information Processing Systems, pp. 4539–4547.
- WU, K.h., LÜ, X.d., Liu, Y.l., 2006. Recent progress of pediatric cardiac surgery in china. Chinese medical journal 119, 2005–2012.
- Wu, L., Li, B., Xia, J., Ji, C., Liang, Z., Ma, Y., Li, S., Wu, Y., Wang, Y., Zhao, Q., 2014. Prevalence of congenital heart defect in guangdong province, 2008-2012. BMC public health 14, 152.
- Yang, F.S.Y., Zhang, L., Xiang, T., Torr, P.H., Hospedales, T.M., 2018. Learning to compare: Relation network for few-shot learning .
- Yang, J., Ren, P., Zhang, D., Chen, D., Wen, F., Li, H., Hua, G., 2017. Neural aggregation network for video face recognition, in: IEEE CVPR, pp. 4362–4371.
- Yang, X.y., LI, X.f., LÜ, X.d., Liu, Y.l., 2009. Incidence of congenital heart disease in beijing, china. Chinese medical journal 122, 1128–1132.
- Zhang, J., Gajjala, S., Agrawal, P., Tison, G.H., Hallock, L.A., Beussink-Nelson, L., Lassen, M.H., Fan, E., Aras, M.A., Jordan, C., et al., 2018. Fully automated echocardiogram interpretation in clinical practice: feasibility and diagnostic accuracy. Circulation 138, 1623–1635.
- Zhang, Y.F., Zeng, X.L., Zhao, E.F., Lu, H.W., 2015. Diagnostic value of fetal echocardiography for congenital heart disease: a systematic review and meta-analysis. Medicine 94.
- Zhao, Q.m., Ma, X.j., Ge, X.l., Liu, F., Yan, W.l., Wu, L., Ye, M., Liang, X.c., Zhang, J., Gao, Y., et al., 2014. Pulse oximetry with clinical assessment to screen for congenital heart disease in neonates in china: a prospective study. The Lancet 384, 747–754.
- Zheng, Y., Barbu, A., Georgescu, B., Scheuering, M., Comaniciu, D., 2008. Four-chamber heart modeling and automatic segmentation for 3-d cardiac ct volumes using marginal space learning and steerable features. IEEE transactions on medical imaging 27, 1668–1681.
- Zhou, B., Andonian, A., Torralba, A., 2018. Temporal relational reasoning in videos. In ECCV .
- Zhou, Z., Huang, Y., Wang, W., Wang, L., Tan, T., 2017. See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification, in: Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on, IEEE. pp. 6776–6785.
- Zou, Y., Yu, Z., Liu, X., Kumar, B., Wang, J., 2019. Confidence regularized self-training. ICCV .