Intelligent Automation & Soft Computing
DOI:10.32604/iasc.2022.023555
Article

Tech Science Press

Classification of Echocardiographic Standard Views Using a Hybrid Attention-based Approach

Zi Ye[1], Yogan Jaya Kumar[2], Goh Ong Sing[2], Fengyan Song[3] and Xianda Ni[4,*]

[1]School of Artificial Intelligence, Wenzhou Polytechnic, Wenzhou, 325035, China
[2]Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, Melaka, 76100, Malaysia
[3]Shanghai Gen Cong Information Technology Co. Ltd., Shanghai, 201300, China
[4]Department of Ultrasonography, The First Affiliated Hospital of Wenzhou Medical University, Wenzhou, 325003, China
*Corresponding Author: Xianda Ni. Email: xianda.ni@gmail.com
Received: 12 September 2021; Accepted: 26 November 2021

Abstract: The determination of the probe viewpoint forms an essential step in automatic echocardiographic image analysis. However, classifying echocardiograms at the video level is complicated, and previous observations concluded that the most significant challenge lies in distinguishing among the various adjacent views. To this end, we propose an ECHO-Attention architecture consisting of two parts. We first design an ECHO-ACTION block, which efficiently encodes Spatio-temporal features, channel-wise features, and motion features. Then, we can insert this block into existing ResNet architectures, combined with a self-attention module to ensure its task-related focus, to form an effective ECHO-Attention network. The experimental results are confirmed on a dataset of 2693 videos acquired from 267 patients that trained cardiologist has manually labeled. Our methods provide a comparable classification performance (overall accuracy of 94.81%) on the entire video sample and achieved significant improvements on the classification of anatomically similar views (precision 88.65% and 81.70% for parasternal short-axis apical view and parasternal short-axis papillary view on 30-frame clips, respectively).

Keywords: Artificial intelligence; attention mechanism; classification; echocardiogram views

1 Introduction

Echocardiography plays a vital role in diagnosing and treating cardiovascular diseases. It is the only imaging method that allows real-time and dynamic observation of the heart and immediate detection of various cardiac abnormalities [1]. However, accurate quantitative evaluation of cardiac structure has been a problem due to the operators' manipulation and the interpretation of echocardiography. For example, there are considerable differences among operators, especially for poor-quality images [2]. It has been proved that the differences between operators and within operators can be reduced with deep learning-based methods. We usually humans subconsciously perform specific steps during each examination. The first critical preprocessing stage is the mode or view classification. Automating this task provides two

leading advantages. Initially, it can foster the organization, storage as well as retrieval of echo images. In the second place, it dominates a key role in the evaluation of the following researches. For instance, measuring the function of a specific valve needs the knowledge of the view beforehand for the reason that different perspectives reveal different valves [3].

Several investigators reported excellent accuracy with utilizing neural network architectures for feature extraction and classification. Representatives of conventional convolutional neural network architectures that have been employed for cardiac view classification contain VGG (Visual Geometry Group), InceptionNet, and ResNet. For example, Zhang et al. developed their previous VGG-like modeling in order to tell from 23 different echo views [4]. The codes and model weights for both works can be obtained via the Internet. In the same way, Madani et al. employed a VGG-based model with the aim to recognize 15 different echo views. The final layer of VGG-16 carries out classification using the Softmax function with the use of 15 nodes [5]. Another cardiac view classification architecture is created on the basis of InceptionNet and put forward to tell from 7 B-mode views [6]. Instead of merely making the extraction of deep characteristics from static images, Gao et al. integrated the output of temporal networks—takes as input the acceleration image, generated by the application of optical flow twice—with that from the spatial stream to classify echocardiogram videos [7]. Recently, we reported on our newly developed view classification architecture driven by action video recognition using a brand-new echocardiogram dataset of the Asian race. In this model, there were 5.7% mislabeled samples on the test set [8].

Current deep learning-based methods for view classification obtained satisfying performance in comparison with the human inter-observer (overall accuracy: around 92%–98% for 7–23 cardiac views). However, upon inspection of the misclassified samples from the literature above, we found that some improvements are needed in the anatomically adjacent imaging planes—like parasternal short-axis views —that are difficult even for experts to judge. Take the papillary muscle level of PSAX (Parasternal Short-Axis) as an example, the left ventricle appears around, and the right ventricular cavity spears semilunar similar to the mitral valve level. The distinctive papillary muscles can only be identified as round structures that bulge into the left ventricular cavity. Furthermore, at the mitral valve level of PSAX, the mitral leaflets are most clearly seen as open nearly to the entire cross-sectional area of the left ventricle in diastole. But during the systole period, the left ventricular cavity gradually shrinks and becoming challenging to observe the fish-mouth mitral valves.

According to the problems mentioned above and limitations, our study aimed to adopt the attention mechanism for designing efficient neural networks. However, admittedly, no other research has applied the attention mechanism to the complex problem of echocardiographic view classification.

We have previously reported preparing and annotating a large patient dataset, covering a range of pathologies and including nine different echocardiographic views, which we used to evaluate our early proposed CNN (convolutional neural network) architectures [8]. In this study, we still use this dataset to design a customized network architecture for the task of echo view classification.

In light of the above, the following presents the leading contributions of the current work:

i) Inspired by the Spatio-temporal, Channel and Motion Excitation (ACTION) module proposed by Wang et al., we design an ECHO-ACTION block that also works in a plug-and-play manner, which can extract appropriate Spatio-temporal patterns, channel-wise features, and motion information to identify the echocardiographic video class. In addition, we adopted a second difference to learning further the motion information inside the network based on the feature level.
ii) The electrocardiogram records heart muscle activity, and differentiating features are most apparent at certain moments. Therefore, our whole ECHO-Attention network employs a self-attention layer after the feature extraction architecture to let the model focus on different phases of a cardiac cycle when making the prediction of the view class. The self-attention module works by comparing

every still image to every other frame in the clips, including itself, and reweighing the feature extraction of each frame to include contextual relevance.

iii) We have demonstrated our proposed ECHO-ACTION module on two different backbones, i.e., ResNet-18 and ResNet-50, and conducted extensive experiments to demonstrate that the proposed method surpasses some advanced neural architectures including CNN+BiLSTM (Bi-directional Long Short-Term Memory) and SpatioTemporal-BiLSTM [8].

2 Attention Mechanism in Deep Learning

Attention mechanism has a substantial and far-reaching impact on deep learning, and it is widely used in various fields. Based on the idea that we need to attend to a specific part of an extensive input (e.g., some words in a sentence of regions in an image) when processing it, the attention mechanism became one of the most powerful concepts. Each element composing the input may have different relevance to the task we are solving: for instance, in machine translation, each word in the source sentence can be more or less relevant for translating the next term; in image captioning, the background regions of an image can be irrelevant to describe an object but crucial to characterize the landscape. To solve this problem, the prevailing solution consists of using attention mechanisms by automatically learning the relevance of any element of the input, i.e., by generating a set of weights (one per element of the input) and take them into account while performing the proposed task.

2.1 Visual Attention Mechanism

Although attention mechanisms were first introduced in NLP (Neuro-Linguistic Programming) for machine translation [9], previous work by Larochelle et al. [10] in Computer Vision had already proposed an object recognition model that learns where to look from scratch using glimpses, inspired by the idea that biological vision systems need to sequentially fixate relevant parts for a specified task because their retina has a limited resolution that falls very quickly with eccentricity. Later, when the work by Mnih et al. [11]—a novel neural network model capable of processing only a selected sequence of regions in an image or video—outperformed state of the art in both static vision tasks (e.g., image classification) and dynamic visual environments (e.g., object tracking), visual attention mechanisms gained popularity.

2.1.1 Hard and Soft

According to the related works mentioned previously, visual attention can be categorized into two classes: hard and soft. Hard attention mechanisms rely on a controller to select the relevant parts of the input, mechanisms like these are not differentiable end-to-end and, for that reason, cannot be trained with the standard backpropagation algorithm. Instead, they require the use of reinforcement learning techniques [12]. Although these models perform well on simple datasets, it has not been easy to use them in real-world applications. To make training easier, soft attention mechanisms commonly used for natural language tasks were introduced in tasks that also require vision (e.g., image captioning and visual question answering). These mechanisms are fully differentiable: they can be plugged in neural networks and trained end-to-end with a gradient backpropagation algorithm [13]. The basic soft-attention network framework is shown in Fig. 1.

2.2 Classification of Video Using Visual Attention

Convolutional Neural Networks have proven to be extremely effective in image classification. The classification of videos rather than images enhances a temporal dimension to the issue of image classification. However, learning temporal dynamics remains a complicated issue. Previous time-series modeling methods have employed LSTMs (Long Short Term Memory networks), optical flow, fused networks as well as hand-crafted features to yield descriptors with both appearance and dynamics

information encoded [14]. Recent work has shown that the advancement of effective attention architectures in computer vision believes the stimulating prospect of finding models possess different and perhaps complementary features to convolutional networks [15]. In this section, we review newly emerging methods by using attention operations developed for video classification applications.

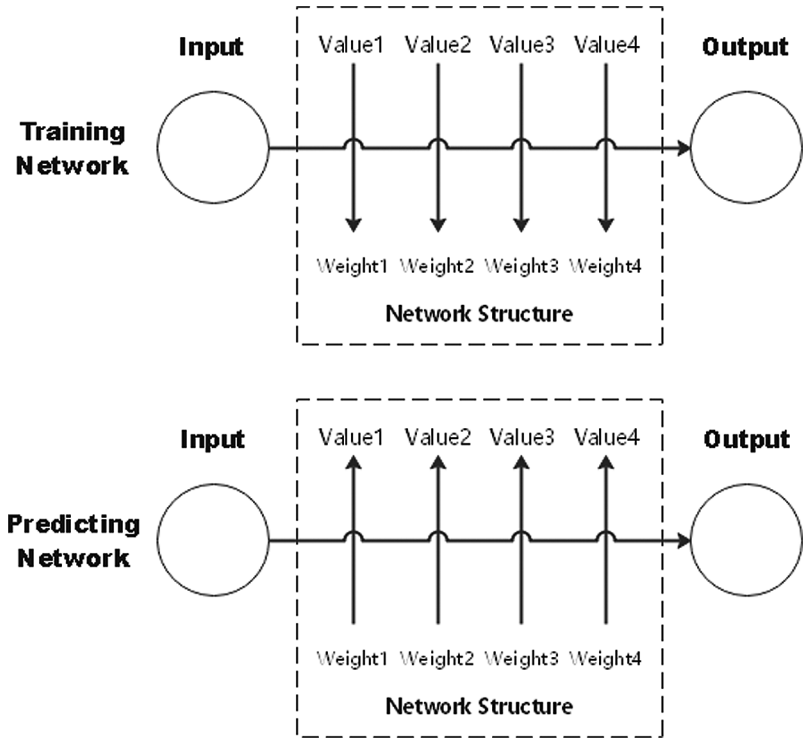

Figure 1: The basic soft-attention network structure

2.2.1 Non-local Neural Networks

This approach is inspired by non-local means operation, which was mostly designed for image denoising. In addition, it also chooses distant pixels to make contributions to the filter response on the basis of the similarity between the patches. By this self-attention design, it computes responses based on long-range dependencies in the image space. Non-local neural networks demonstrated that the core attention mechanism in Transformers could produce good results on video tasks. However, it is confined to processing only short clips [16].

2.2.2 LSTMs with Attention

Sharma et al. put forward a soft-attention-based model for action recognition in videos, learning which parts in the frames are related to the task at hand and lays stronger emphasis on them as well as categorizes videos after taking a few glimpses [14]. However, the soft-attention models proposed using weighted averages are even computationally costly since they expect all the features to conduct dynamic pooling.

2.2.3 Video Transformer Network

Neimark et al. propose Video Transformer Network (VTN) that first obtain frame-wise features using 2D CNN and apply a Transformer encoder (Longformer) on top to learn temporal relationships. Longformer is an attractive choice to process sequences with an arbitrary length n due to its complexity, making VTN particularly suitable for modeling long videos where interactions between entities are spread throughout

the video length. The classification token [CLS] is passed through a fully connected layer to recognize actions or events [17].

2.2.4 Action-net

Spatial-temporal, channel-wise, and motion patterns are regarded as three complementary and crucial types of information for video recognition. Therefore, Wang et al. introduced a novel ACTION module composing of the following three paths, respectively, Spatio-Temporal Excitation (STE) path, Channel Excitation (CE) path, and Motion Excitation (ME) path. The experiments on action recognition datasets with various backbones show competitive performance [18].

3 Approach

The current section presents technical details for our proposed network architecture. Firstly we describe an ECHO-ACTION block that utilizes multipath excitation for Spatio-temporal features, channel-wise features, and motion features of the cardiac cycle activity. This block can be inserted into existing ResNet architecture (here we demonstrate on ResNet-18 and ResNet-50) to form our ECHO-Attention model. Afterward, a self-attention module is added after the feature extraction to ensure the model focuses on task-related time steps.

3.1 Spatio-Temporal Module and Channel Module

The Spatio-Temporal module (STM) and Channel module (CM), as both shown in Fig. 2, are designed very similarly to Spatio-Temporal Excitation (STE) and Channel Excitation (CE) in ACTION-Net [18]. The main difference is that our input tensor is 4D, i.e., (N × T, C, H, W). For STM, it generates a Spatio-temporal mask M used for element-wise multiplying the input X across all channels. The final output can be interpreted as Eq. (1). For CM, another channel mask M is obtained, and its output is formulated as the same as in Eq. (1) using the newly generated mask.

$$Y = X + X \odot M \tag{1}$$

3.2 Motion Module

Our Motion module (MM) is inspired by Motion Excitation (ME) previously [18], which aims to calculate feature-level temporal differences. However, unlike previous work that proposed a block form extracting only adjacent frame motions, we use the MM to integrate further with the second difference to generate a stronger motion information.

As illustrated in Fig. 3, firstly, the motion information is modeled by the two successive frames. Then, we adopt the squeeze strategy by using 1 × 1 2D convolutional layers and batch normalization, referred to as Eqs. (2) and (3).

$$X_r = K_1 \times X \tag{2}$$

$$X_b = batchnorm(X_r) \tag{3}$$

where K_1 is a 1 × 1 2D convolutional layer, and we squeeze the number of channels by a scale ratio of 16 in this work.

Given the feature X_b, we processed splitting along the time direction to get $X_b = [X_b(1), \ldots, X_b(t)]$, and meanwhile, the motion feature X_b is modeled the following operations, represented as the Eqs. (4) and (5).

$$X_c = K \times X_b \tag{4}$$

$$X_c = [X_c(1), \ldots, X_c(t)] \tag{5}$$

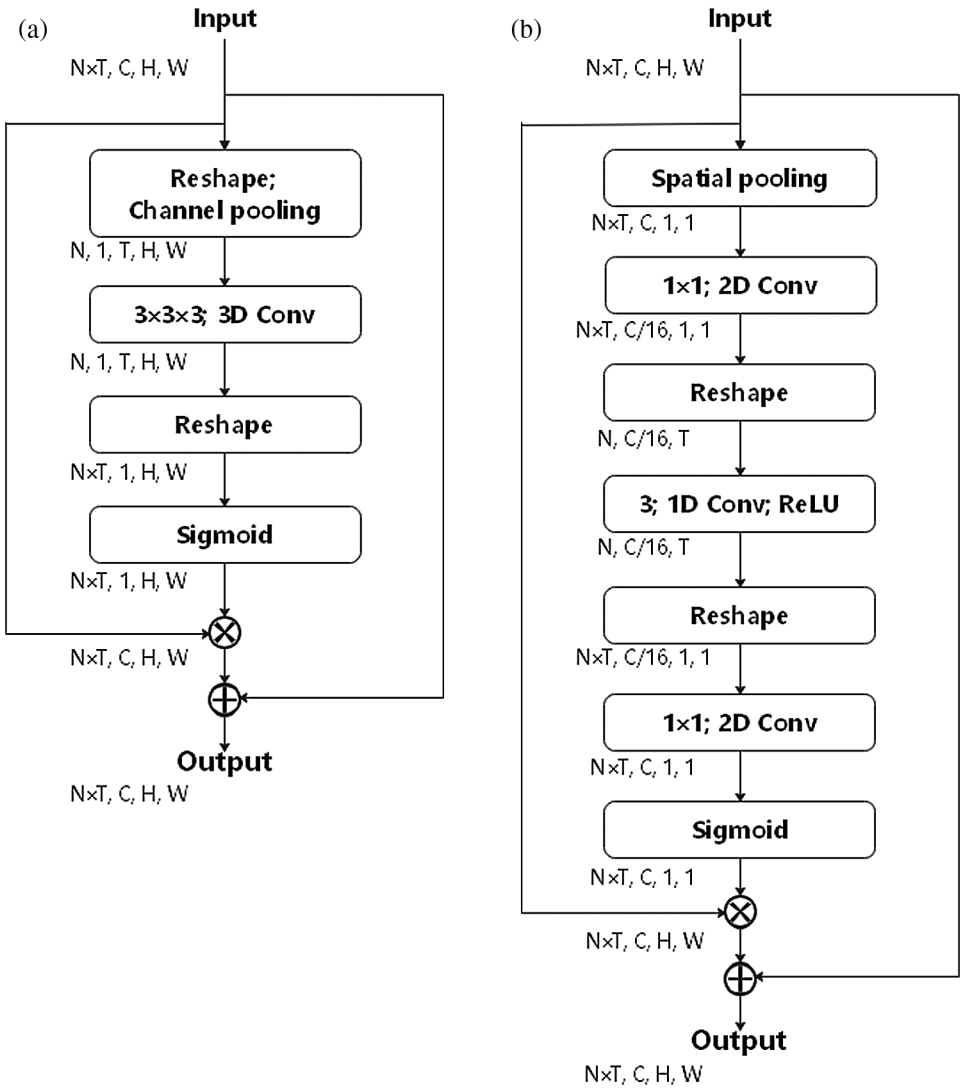

Figure 2: (a) Spatio-temporal module (STM), (b) Channel module (CM)

The first difference of motion information takes place following, which can be formulated as Eq. (6). According to the temporal dimension, the motion feature is then concatenated to each other, and zero is padded to the first element. The X_d is then processed by spatial average pooling, and the feature output and the mask can finally be achieved similarly as in Eq. (1).

$$X_d = X_c(t+1) - X_b(t) \tag{6}$$

In order to obtain a more powerful motion information generator, we decided to apply the second difference method. To calculate the second difference, the given motion feature X_d is reshaped and fed to a 2D convolutional layer with kernel size 3×3 to get X_e. We then select 2 consecutive feature values on the time domain and process subtraction to get the second difference values, which can be represented as the Eq. (7), followed by zero-padding to the first two elements, average spatial pooling, 1×1 2D convolutional layer, and sigmoid to get attention map, as described before for the first difference method.

$$X_f = X_e(t+1) - X_e(t) \tag{7}$$

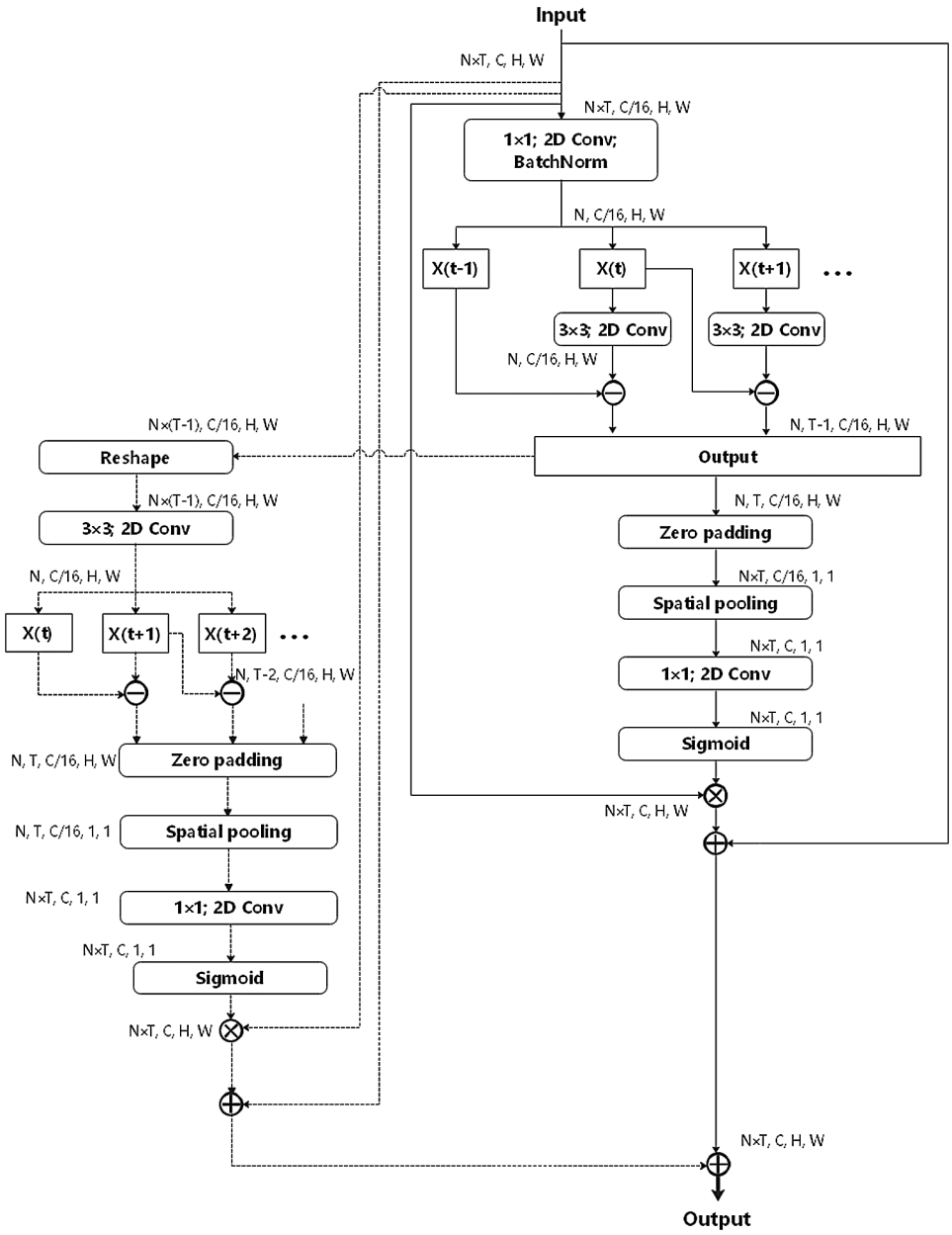

Figure 3: Motion module (MM)

3.3 ECHO-ACTION Block

The overall ECHO-ACTION module takes the element-wise addition of three excited features generated by STM, CM, and MM, respectively (see ECHO-ACTION block in Fig. 4). By doing this, the output of the ECHO-ACTION block can perceive information from a Spatio-temporal perspective, channel interdependencies, and motion. Therefore, multi-type information in videos can be activated through this hybrid attention mechanism. Due to the fact that the ECHO-ACTION block proposed in this study is consistent with the ordinary residual block, it is not difficult for us to incorporate it into any current ResNet architecture in order to construct our new network. The following section will focus on the general architecture of the ECHO-Attention network.

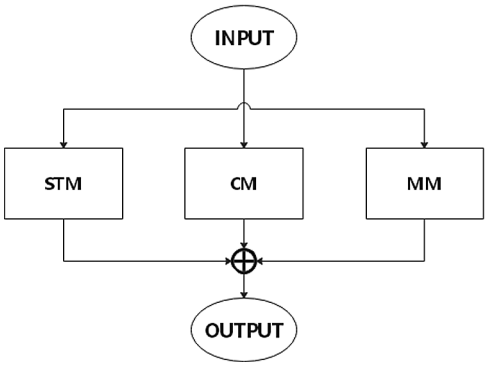

Figure 4: ECHO-ACTION block

3.4 Architecture

Our proposed ECHO-Attention network architecture for ResNet-50 is shown in Fig. 5, and it mainly consists of two parts. The echo frames as input first pass through the ResNet-50 backbone together with the ECHO-ACTION block, wherein the ECHO-ACTION block is inserted at the beginning of each residual block. And it does not require any modification for the original components in the block. This partial framework is used to generate the multi-type information by effectively adopting multipath excitation.

A multi-head self-attention mechanism is then adopted to make the network discover the feature vectors that should receive more attention. Here the self-attention sublayers employ 8 attention heads. The results from each head are concatenated to form the sublayer output, and a parameterized linear transformation is applied [19].

The last encoding part maps the input video frame to a tensor with dimension (N × 30, 512), which are treated as an input sequence that each attention head operates on. The input $x = (x_1, \ldots, x_{30})$ of 30 elements in which x_i is a 512-D vector is used to compute a new sequence $z = (z_1, \ldots, z_{30})$ of the same length in which z_i is also a 512-D vector.

Each output element, z_i, is computed as a weighted sum of linearly transformed input elements, shown in Eq. (8).

$$z_i = \sum_{j=1}^{30} a_{ij}\left(x_j W^V\right) \tag{8}$$

Each weight coefficient, a_{ij}, is computed using a SoftMax function as Eq. (9).

$$a_{ij} = \frac{exp\left(e_{ij}\right)}{\sum_{k=1}^{30} exp(e_{ik})} \tag{9}$$

And e_{ij} is computed using a compatibility function that compares two input elements, shown in Eq. (10).

$$e_{ij} = \frac{\left(x_i W^Q\right)\left(x_j W^K\right)^T}{\sqrt{d_z}} \tag{10}$$

A scaled dot product was chosen for the compatibility function, which enables efficient computation. In addition, linear transformation of the inputs adds sufficient expressive power [20]. W^Q, W^K, W^V are parameter matrices. The above parameter matrices have special per layer and attention head.

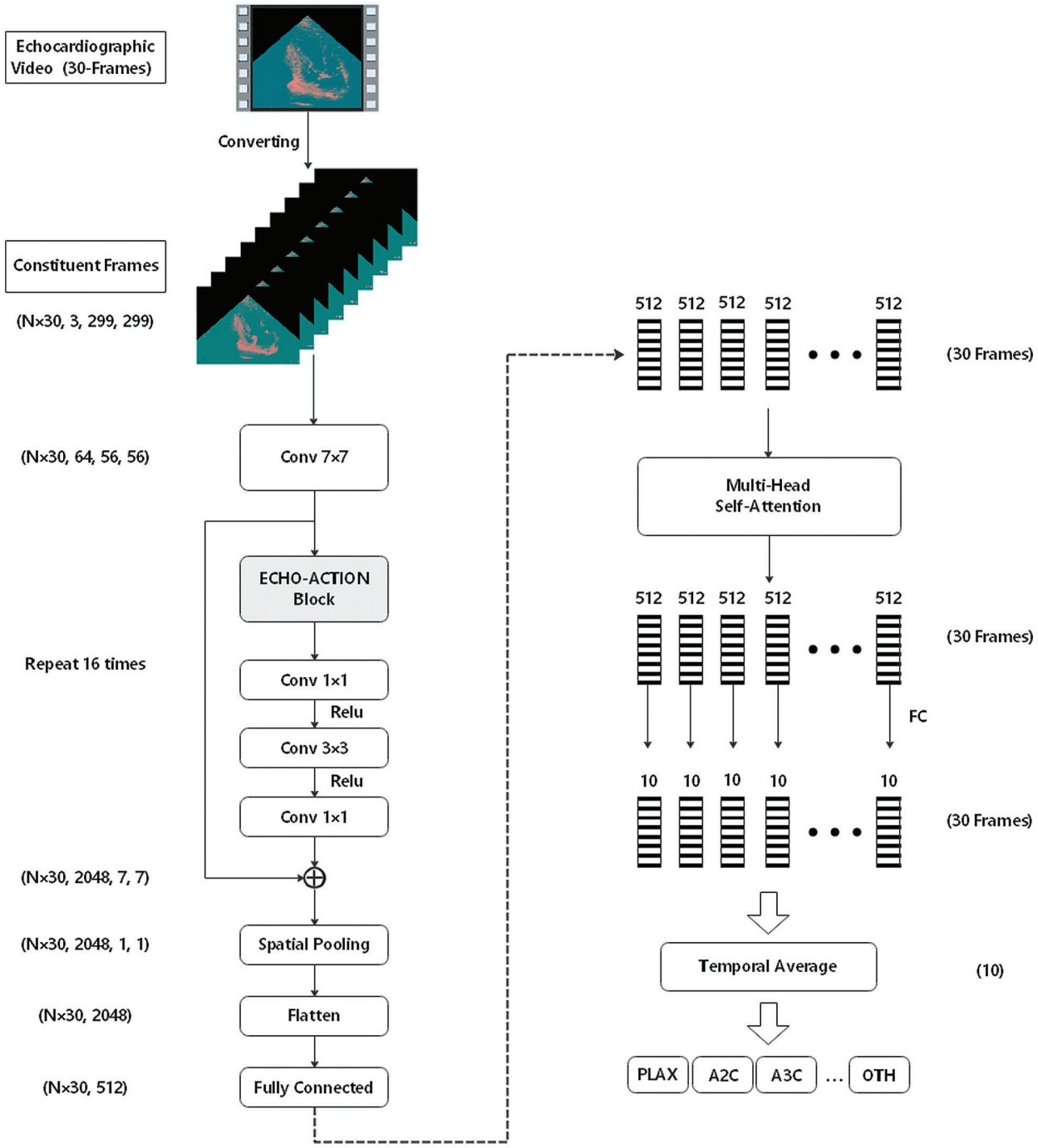

Figure 5: ECHO-Attention for ResNet-50 architecture

Each element of the obtained new sequence $z = (z_1, \ldots, z_{30})$ is finally mapped to a fully connected layer of ten nodes, corresponding to ten probabilities for each processed image. The arithmetic mean of categories determines the final predicted label over these 30 consecutive frames.

4 Experiments

Pytorch was utilized to implement the models. For the computationally intensive stage of video analysis, a GPU (Graphics Processing Unit) server equipped with two NVIDIA A100 with 40 GB of memory was rented. Additionally, each GPU has a mini-batch of 4 video clips. Here, a detailed description of experimental implementation is provided.

4.1 Datasets Introduction

In this section, a brief account of the patient dataset used is provided. Our study discusses two main imaging windows during a standard echocardiographic examination: the parasternal and apical windows. First, the parasternal and apical windows can be acquired with the patient positioned in the left lateral decubitus position, considering that the patient is capable of assuming this position. Then a sequential series of images are obtained in each window and used to assess the cardiac functions from different perspectives. Here we aimed at including subclasses of given echocardiographic views, which are outlined in Fig. 6. In general, the more numerous the view classes, the more complex the task of differentiating the views for the deep learning model.

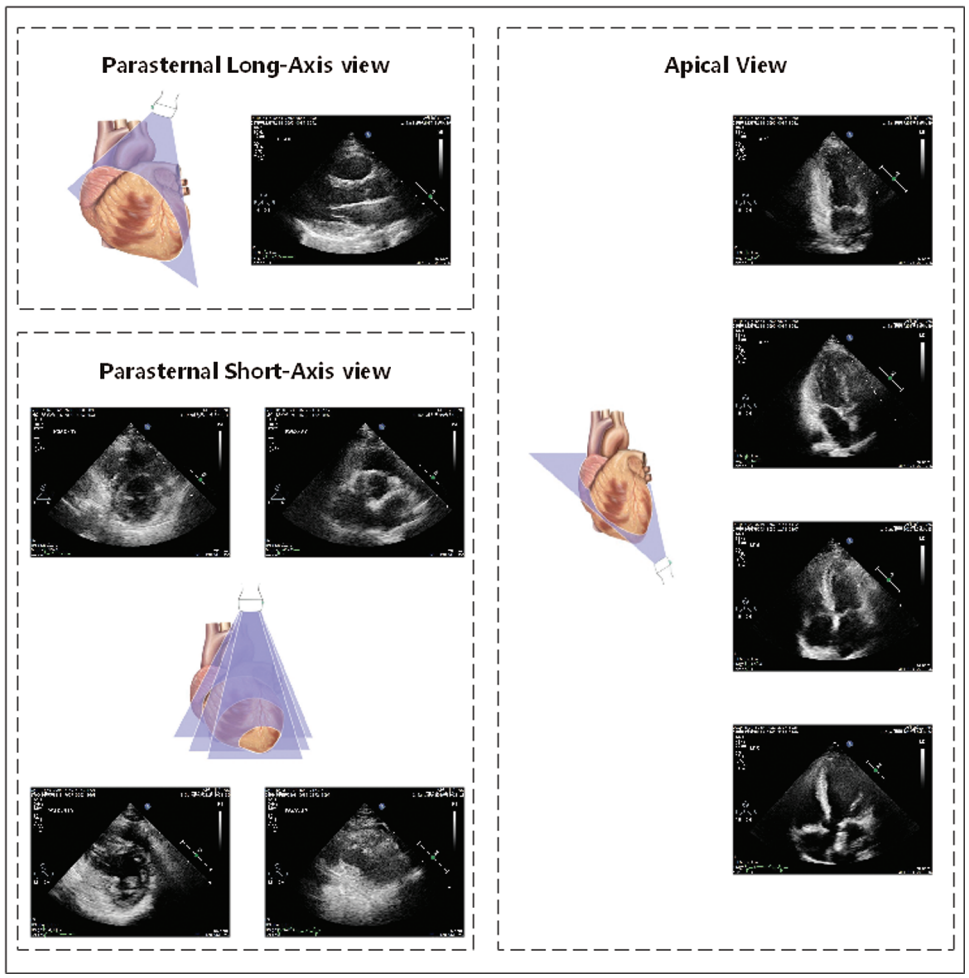

Figure 6: Inclusion of 9 different anatomical echocardiographic views: apical two-chamber (A2C), apical three-chamber (A3C), apical four-chamber (A4C), apical five-chamber (A5C), parasternal long-axis (PLAX), parasternal short-axis aortic valve focused (PSAX-AV), parasternal short-axis apical focused (PSAX-AP), parasternal short-axis mitral valve focused (PSAX-MV), and parasternal short-axis papillary focused (PSAX-MID); A class of "others" is also investigated for other cardiac aspects

All the involved datasets were acquired and de-identified, with waived consent in accordance with the Institutional Review Board (IRB) at a private hospital in Malaysia. The acquisition of the images was conducted by experienced echocardiographers and, based on the standard regulations and guidelines with the use of ultrasound equipment from Philips manufacturer. Only studies with whole patient demographic data and without intravenous contrast administration were covered.

In DICOM (Digital Imaging and Communications in Medicine) format, random echocardiogram studies of 267 patients and their associated video loops, together 2693, were extracted from the current hospital's echocardiogram database. Videos considered to show no identifiable echocardiographic features or which depicted more than one view were excluded. Among these, 2443 (90.7%) videos with the following classes are selected and annotated manually by a board-certified echocardiographer, finally categorized into nine different folders: PLAX, PSAX-AV, PSAX-MV, PSAX-AP, PSAX-MID, A4C, A5C, A3C, and A2C, and along with the "OTHERS" to put the remaining 250 videos because usually, a comprehensive study comprises more required views and measurements, such as Suprasternal (SSN) and Subcostal (SC). The relative distribution of echo view classes labeled by expert cardiologists is displayed in Fig. 7.

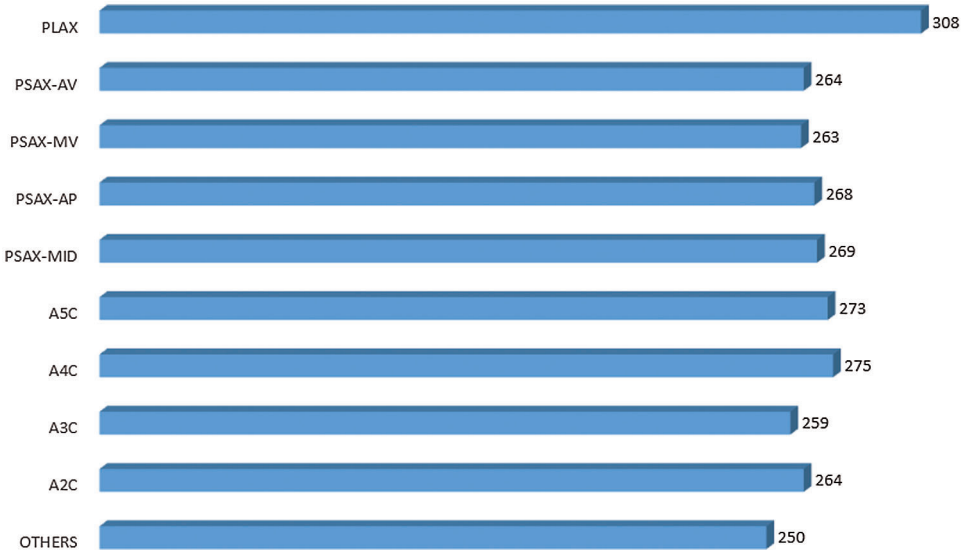

Figure 7: Distribution of numbers of video samples in a given class

4.2 Preprocessing

Each DICOM-formatted echocardiogram video comes with a set of interfering visual features and is less relevant for view identification, such as class labels, patient digital identifiers, study duration, etc. Therefore, we employed field of view segmentation proposed in our previous work [8] as preprocessing to localize relevant visual features and also simplify the classification task.

4.3 Implementation Details

Fig. 8 depicts the overview of the training procedure for this study. After prior processing, the dataset was randomly divided into training, validation, and testing sub-datasets in a 2.5:1:1 ratio. Each sub-datasets contained clips from separate patient studies to maintain sample independence. Then each complete sample video is cut into several 30-frame short clips with an interval of 5 frames, and the datasets descriptions are given in Tab. 1.

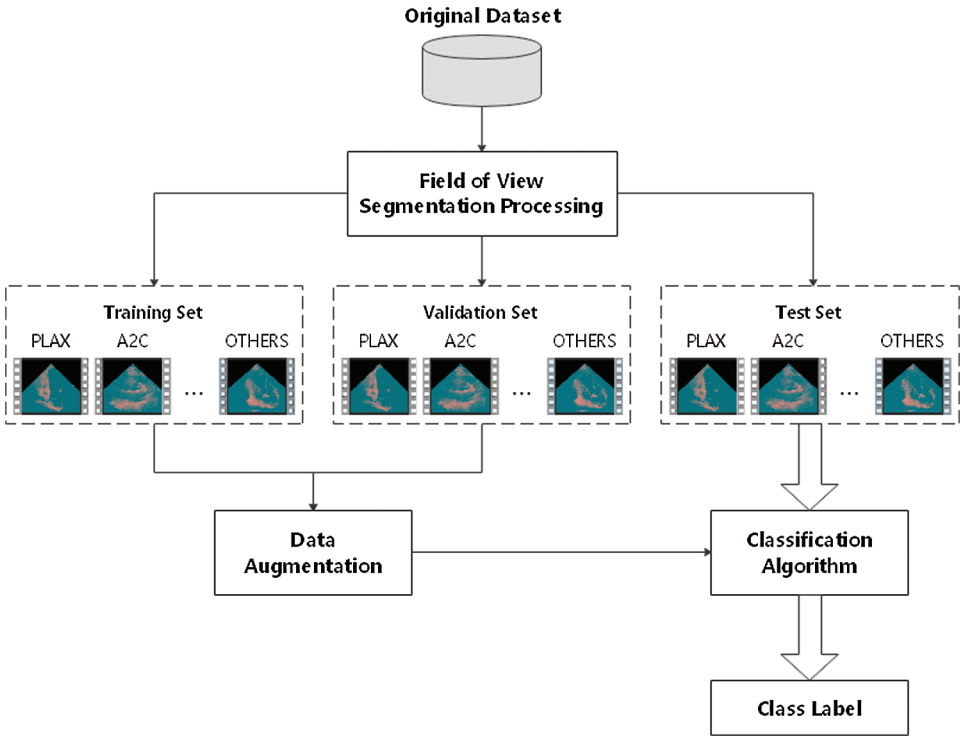

Figure 8: The training procedure for this study

Table 1: The overview of the datasets of 30-frame

	PLAX	PSAX-AV	PSAX-MV	PSAX-AP	PSAX-MID	A5C	A4C	A3C	A2C	OTHERS
Train	2310	2110	2081	2155	2244	1883	1942	1937	1877	2008
Valid	922	827	913	889	836	776	755	694	739	846
Test	949	792	796	810	835	674	694	663	687	925

All the 30-frame echocardiographic segments are converted to consecutive 30 frames for the following training processing. The original resolution of each frame is 600 × 800. Meanwhile, a series of data argumentation strategies—a random crop with a scale of 0.75 to 1, rotations of up to 20 degrees, and horizontal/vertical flips—are applied. Each cropped frame was finally resized to 299 × 299, which was used for training the model. Thus, the input fed to the model is of the size 8 × 30 × 3 × 299 × 299, in which 8 is the batch size.

Our proposed network is trained in the following three phases: 1) Pre-training the ResNet-50 model with the echocardiographic sample dataset for 20 epochs; 2) Pre-training the ResNet-50 backbone and the ECHO-ACTION block for 10 epochs. In this stage, the weights saved from the last step were applied as the initialization for the ResNet-50 part; 3) Training the parameters of the entire network for 10 epochs. Like stage 2, we save the weights learned before, and the dropout of probability 0.3 was added to the self-attention module. The learning rates of these three stages were set to $1e^{-3}$, $1e^{-4}$, and $1e^{-5}$ separately, and an ADAM (Adaptive Momentum Estimation) optimizer with ReduceLROnPlateau scheduler was used.

It is crucial to illustrate the loss function we used for stage 3 mentioned above. Let l_i be the loss obtained for each sample, and the computation is shown as Eq. (11).

$$l_i = CrossEntropy\ (Pred_i,\ Label_i) \tag{11}$$

Thus, the loss for all the samples in one minibatch is Eq. (12).

$$L = \{l_1, l_2,\ l_3, \cdots, l_n\} \tag{12}$$

where n is the batch size.

Then we sorted the loss set in descending order, and we get Eq. (13).

$$L' = \left\{l_1{}', l_2{}',\ l_3{}', \cdots, l_n{}'\right\} \tag{13}$$

We need to find K that satisfying both Eqs. (14) and (15).

$$\sum_{i=1}^{k} l_i' \geq \frac{1}{2}\sum_{i=1}^{n} l_i' \tag{14}$$

$$\sum_{i=1}^{k-1} l_i' < \frac{1}{2}\sum_{i=1}^{n} l_i' \tag{15}$$

The backpropagation algorithm updates the neural network's weights by minimizing the loss function Eq. (16).

$$Loss = \sum_{i=1}^{k} l_i' \tag{16}$$

4.4 Classification Evaluation Metrics

Performance metrics are used to evaluate and for checking the quality of performance by the algorithms. Accuracy is considered to be one of the most widely applied performance metrics in classification. Classification accuracy denotes the number of instances (such as images or pixels) that are properly categorized by the total number of instances in the dataset ($(TP + TN)/(TP + TN + FP + FN)$), in which TP, TN, FP, and FN stand for true positive, true negative, false positive, and false negative, respectively. Classification metrics are derived from the confusion matrix, revealing the number of correct and incorrect classifications when compared with the ground truth labels. In addition, this study also calculates precision ($TP/(TP + FP)$) and recalls ($TP/(TP + FN)$) on the basis of the confusion matrix.

5 Results and Discussions

The current section compares our approach with the state-of-the-art action recognition architectures, including CNN+BiLSTM and SpatioTemporal-BiLSTM. As illustrated in Tab. 2, the ECHO-Attention model consistently outperforms these two baseline models. The results of the baseline method CNN +BiLSTM remain comparatively low because it only contains the component that is able to model the Spatio-temporal information. Whereas for the SpatioTemporal-BiLSTM, the optical flow employed only denotes the motion features between the neighboring frames. As a result, this flow stream is short of the capability of capturing the long-range temporal relationship. By further employing a second difference operation to the part of the Motion Module, the ECHO-ACTION block takes additional motion modeling to the network, which significantly improves the ECHO-Attention network compared to both baselines.

Table 2: Table demonstrating the overall accuracy for each of the networks on the datasets of 30-frame

	ResNet-18	ResNet-50
CNN+BiLSTM	91.74%	91.83%
SpatioTemporal-BiLSTM	92.19%	92.12%
ECHO-Attention	93.25%	93.85%

It is also noting that Tab. 2 demonstrates classification performance for the ECHO-Attention model employing two backbones, i.e., ResNet-18 and ResNet-50. It can be noticed that ResNet-50 outperforms ResNet-18 regarding the accuracy for the ECHO-Attention architecture, which indicates that ResNet-50 benefits mainly from the ECHO-ACTION block.

Furthermore, we investigate the design of our ECHO-Attention network (ResNet-50) concerning the classification accuracy performance on the entire echo videos. First, the complete testing video sample is split into several 30-frame videos. Then, the individual testing short clip is predicted by our proposed models and classified by referring to the most possible view. Finally, the plurality voting of multiple 30-frame videos generated from the entire video is used to classify the test videos.

The superiority of the ECHO-Attention on the entire echo video is also quite impressive. From Tab. 3, it can be noticed that the overall accuracy of the ECHO-Attention network is improved by 0.51% and 1.01% compared to Xception+BiLSTM and SpatioTemporal-BiLSTM, which have been conducted in our last project and shown to be effective [8].

Table 3: Classification accuracy on entire videos

Architectures	Overall accuracy
Xception+BiLSTM	94.30%
SpatioTemporal-BiLSTM (Xception)	93.80%
Resnet50+BiLSTM	93.13%
SpatioTemporal-BiLSTM (Resnet50)	93.13%
ECHO-Attention (ResNet-50)	94.81%

The confusion matrix, precision values, and recall values of the ECHO-Attention network for ResNet-18 together with ResNet-50, evaluated on the datasets of 30-frame segments, are provided in Fig. 9 and Tab. 4. As a result, echo views with distinct characteristics are more accessible for the model to distinguish. For instance, the parasternal long-axis view seems to have higher correct identification rates, and the network is confused only on four occasions between PLAX and PSAX-MV.

According to the literature review, previous studies found that errors occur predominantly clustered between particular views, denoting anatomically adjacent imaging planes. The PSAX-MID view proves to be the hardest one to detect, as the classifier usually is confused between this view with other parasternal short-axis views, such as PSAX-AP. The reason for these views the model found most challenging to differentiate correctly is that distinctive features are merely partly in view or only in view during part of the cardiac cycle. Interestingly, these views are similar in appearance to human eyes, even for cardiologist experts.

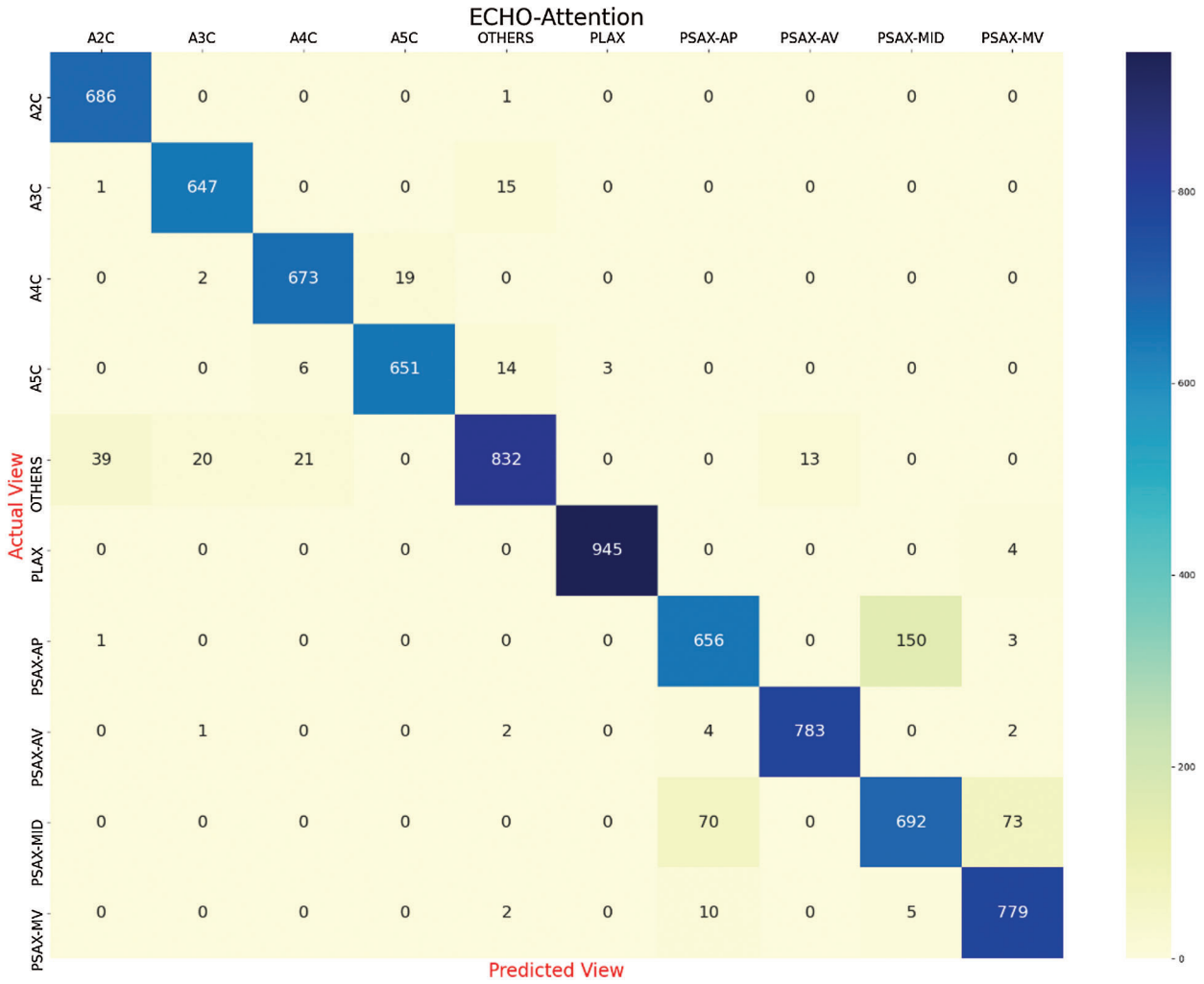

Figure 9: Confusion matrices for echo view classification. The ECHO-Attention network (ResNet-50) was evaluated on the datasets of 30-frame segments

According to Section 3.4, the self-attention module is supplemented to our ECHO-Attention network in order to guarantee its concentration on task-related time steps. It allows the individual frame along the time direction to interact with each other, so the frames with the view-defining structures can be most weighted and contributed mainly to the view classification. The attention weights for predicting different cardiac views are obtained by focusing on different phases of a cardiac cycle throughout the attention layer. As listed in Tabs. 4 and 5, Our ECHO-Attention network significantly decreases similar views' misclassification compared with other successful architectures. We report higher precision and recall values of PSAX-MID and PSAX-AP, gaining 7.67% and 9.32% precision improvement, respectively, compared with SpatioTemporal-BiLSTM (ResNet-50).

This is also evident in Tab. 6 and Fig. 10, which display the precision, recall, and confusion matrix of the ECHO-Attention network (ResNet-50) testing on the datasets of entire videos. Noticeably, the results also demonstrate the importance of motion modeling for echocardiographic datasets and the attention corresponding to the images showing the most view-defining structures.

Table 4: Precision and recall results using ECHO-Attention network (ResNet-50) on the datasets of 30-frame segments

Views	Precision	Views	Recall
A2C	94.36%	A2C	99.85%
A3C	96.57%	A3C	97.59%
A4C	96.14%	A4C	96.97%
A5C	97.16%	A5C	96.59%
OTHERS	96.07%	OTHERS	89.95%
PLAX	99.68%	PLAX	99.58%
PSAX-AP	88.65%	PSAX-AP	80.99%
PSAX-AV	98.37%	PSAX-AV	98.86%
PSAX-MID	81.70%	PSAX-MID	82.87%
PSAX-MV	90.48%	PSAX-MV	97.86%

Table 5: Precision and recall results using SpatioTemporal-BiLSTM (ResNet-50) on the datasets of 30-frame segments

Views	Precision	Views	Recall
A2C	91.86%	A2C	98.54%
A3C	99.08%	A3C	97.29%
A4C	94.88%	A4C	98.85%
A5C	99.19%	A5C	90.65%
OTHERS	95.66%	OTHERS	92.86%
PLAX	100%	PLAX	99.37%
PSAX-AP	79.33%	PSAX-AP	79.14%
PSAX-AV	97.64%	PSAX-AV	99.12%
PSAX-MID	74.03%	PSAX-MID	77.49%
PSAX-MV	92.61%	PSAX-MV	89.70%

Table 6: Precision and recall results using ECHO-Attention network (ResNet-50) on the datasets of entire videos

Views	Precision	Views	Recall
A2C	96.67%	A2C	100%
A3C	98.21%	A3C	98.21%
A4C	98.31%	A4C	98.31%
A5C	98.31%	A5C	98.31%
OTHERS	96.30%	OTHERS	91.23%
PLAX	100%	PLAX	100%

(Continued)

Table 6 (continued)

Views	Precision	Views	Recall
PSAX-AP	90.38%	PSAX-AP	78.33%
PSAX-AV	98.28%	PSAX-AV	98.28%
PSAX-MID	79.69%	PSAX-MID	85.00%
PSAX-MV	92.06%	PSAX-MV	100%

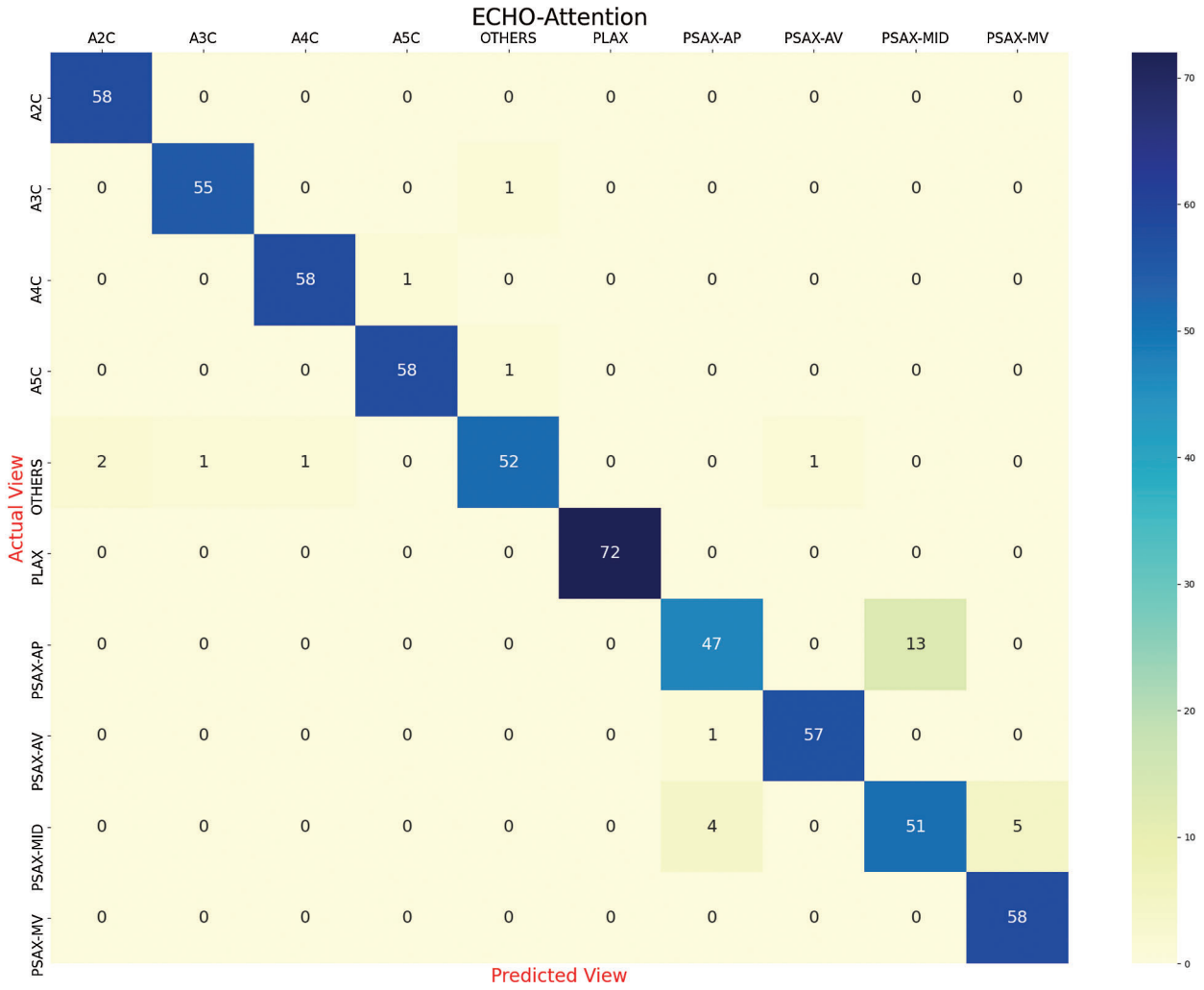

Figure 10: Confusion matrices for echo view classification. The ECHO-Attention network (ResNet-50) was evaluated on the datasets of entire videos

6 Conclusions

To conclude, there exists significant interest in AI (Artificial Intelligence) systems that can support a cardiologist in diagnosing echocardiograms. In the meanwhile, automatic echo view classification is the first step. In this study, we presented a simple and effective architecture called ECHO-Attention, which is

used for the automated identification of 9 different anatomical echocardiographic views (in addition to a class of "others") in a dataset of 2693 videos acquired from 267 patients. We first target designing an ECHO-ACTION block that utilizes multipath excitation for Spatio-temporal features, channel-wise features, and motion features. Any ResNet architecture could leverage this proposed block. Also, the afterward self-attention module helps the network focus on most corresponding and related segments and gives a better prediction. According to the obtained results, the method proposed in this study (ECHO-Attention network for ResNet-50) can achieve comparable classification performance. Such a model can thus be used for real-time detection of the classic echo view.

Acknowledgement: This work is supported by Pantai Hospital Ayer Keroh, Malaysia, and the authors would also like to thank Universiti Teknikal Malaysia Melaka for supporting this research.

Funding Statement: This work was supported in part by the Research Project of Wenzhou Polytechnic, China, under Grant WZY2021011.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] R. Erbel, W. Daniel, C. Visser, R. Engberding, J. Roelandt *et al.,* "Echocardiography in diagnosis of aortic dissection," *Lancet*, vol. 333, no. 8636, pp. 457–461, 1989.

[2] R. Hoffmann, H. Lethen, T. Marwick, M. Arnese, P. Fioretti *et al.,* "Analysis of interinstitutional observer agreement in interpretation of dobutamine stress echocardiograms," *Journal of the American College of Cardiology*, vol. 27, no. 2, pp. 330–336, 1996.

[3] G. Zamzmi, L. Y. Hsu, W. Li, V. Sachdev and S. Antani, "Harnessing machine intelligence in automatic echocardiogram analysis: Current status, limitations, and future directions," *IEEE Reviews in Biomedical Engineering*, vol. 14, pp. 181–203, 2021.

[4] J. Zhang, S. Gajjala, P. Agrawal, G. H. Tison, L. A. Hallock *et al.,* "Fully automated echocardiogram interpretation in clinical practice," *Circulation*, vol. 138, no. 6, pp. 1623–1635, 2018.

[5] A. Madani, R. Arnaout, M. Mofrad and R. Arnaout, "Fast and accurate view classification of echocardiograms using deep learning," *NPJ Digital Medicine*, vol. 1, no. 1, pp. 1–8, 2018.

[6] A. Østvik, E. Smistad, S. A. Aase, B. O. Haugen and L. Lovstakken, "Real-time standard view classification in transthoracic echocardiography using convolutional neural networks," *Ultrasound in Medicine and Biology*, vol. 45, no. 2, pp. 374–384, 2019.

[7] X. Gao, W. Li, M. Loomes and L. Wang, "A fused deep learning architecture for viewpoint classification of echocardiography," *Information Fusion*, vol. 36, no. 2, pp. 103–113, 2017.

[8] Z. Ye, Y. J. Kumar, G. O. Sing, F. Song, X. Ni *et al.,* "Artificial intelligence-based echocardiogram video classification by aggregating dynamic information," *KSII Transactions on Internet & Information Systems*, vol. 15, no. 2, pp. 500–521, 2021.

[9] D. Bahdanau, K. H. Cho and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. Int. Conf. on Learning Representations*, San Diego, SD, USA, 2015.

[10] H. Larochelle and G. Hinton, "Learning to combine foveal glimpses with a third-order Boltzmann machine," in *Proc. Neural Information Processing Systems*, Vancouver, BC, Canada, pp. 1243–1251, 2010.

[11] V. Mnih, N. Heess and A. Graves, "Recurrent models of visual attention," in *Proc. Neural Information Processing Systems*, Palais des Congrès de Montréal, Montréal, Canada, 2014.

[12] A. C. Schütz, D. I. Braun and K. R. Gegenfurtner, "Eye movements and perception: A selective review," *Journal of Vision*, vol. 11, no. 5, pp. 9, 2011.

[13] S. Chaudhari, V. Mithal, G. Polatkan and R. Ramanath, "An attentive survey of attention models," *ACM Transactions on Intelligent Systems and Technology (TIST),* vol. 12, no. 5, pp. 1–32, 2019.

[14] S. Sharma, R. Kiros and R. Salakhutdinov, "Action recognition using visual attention," in *Proc. Int. Conf. on Learning Representations*, San Juan, Puerto Rico, 2016.

[15] R. Ranftl, A. Bochkovskiy and V. Koltun, "Vision transformers for dense prediction," in *Proc. IEEE/CVF Int. Conf. on Computer Vision*, Montreal, Canada, pp. 12179–12188, 2021.

[16] X. Wang, R. Girshick, A. Gupta and K. He, "Non-local neural networks," in *Proc. Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 7794–7803, 2018.

[17] D. Neimark, O. Bar, M. Zohar and D. Asselmann, "Video transformer network," *arXiv preprint arXiv:2102.00719*, 2021.

[18] Z. Wang, Q. She and A. Smolic, "ACTION-Net: Multipath excitation for action recognition," in *Proc. Computer Vision and Pattern Recognition*, Nashville, USA, pp. 13214–13223, 2021.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones *et al.,* "Attention is all you need," in *Proc. Neural Information Processing Systems*, Long Beach California, USA, pp. 6000–6010, 2017.

[20] P. Shaw, J. Uszkoreit and A. Vaswani, "Self-attention with relative position representations," in *Proc. Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, New Orleans, Louisiana, USA, pp. 464–468, 2018.