Transformer-Based Spatio-Temporal Analysis for Classification of Aortic Stenosis Severity From Echocardiography Cine Series

N. Ahmadi, M. Y. Tsang, A. N. Gu, T. S. M. Tsang, and P. Abolmaesumi, *Senior Member, IEEE*

***Abstract*—Aortic stenosis (AS) is characterized by restricted motion and calcification of the aortic valve and is the deadliest valvular cardiac disease. Assessment of AS severity is typically done by expert cardiologists using Doppler measurements of valvular flow from echocardiography. However, this limits the assessment of AS to hospitals staffed with experts to provide comprehensive echocardiography service. As accurate Doppler acquisition requires significant clinical training, in this paper, we present a deep learning framework to determine the feasibility of AS detection and severity classification based only on two-dimensional echocardiographic data. We demonstrate that our proposed spatio-temporal architecture effectively and efficiently combines both anatomical features and motion of the aortic valve for AS severity classification. Our model can process cardiac echo cine series of varying length and can identify, without explicit supervision, the frames that are most informative towards the AS diagnosis. We present an empirical study on how the model learns phases of the heart cycle without any supervision and frame-level annotations. Our architecture outperforms state-of-the-art results on a private and a public dataset, achieving 95.2% and 91.5% in AS detection, and 78.1% and 83.8% in AS severity classification on the private and public datasets, respectively. Notably, due to the lack of a large public video dataset for AS, we made slight adjustments to our architecture for the public dataset. Furthermore, our method addresses common problems in training deep networks with clinical ultrasound data, such as a low signal-to-noise ratio and frequently uninformative frames. Our source code is available at: https://github.com/neda77aa/FTC.git**

***Index Terms*—Aortic stenosis, cardiac imaging, spatio-temporal analysis, temporal localization, ultrasound.**

I. Introduction

AORTIC stenosis (AS) [1] is a severe valvular heart disease associated with thickening and calcification of aortic valve (AV) leaflets. This restricts the motion of AV leaflets and reduces blood flow from the left ventricle to the rest of the body. AS becomes more prevalent with age, making the problem more significant alongside an aging demographic. Clinically significant AS is fatal, with a 5-year mortality rate of 56% and 67% for those classified with moderate and severe AS, respectively, if left untreated [2]. Thus, an accessible method of screening is essential for early detection and timely intervention of AS.

Echocardiography (echo) is the current clinical standard for determining the severity of AS, where three clinical markers (AV area, peak velocity of the valvular jet and mean pressure gradient) are determined primarily based on Doppler measurements [3]. This information is interpreted by experienced cardiologists based on the clinical guidelines to make a diagnosis. However, Doppler imaging is technically challenging for less experienced users, resulting in high interobserver variability for AS diagnosis.

Recently, a body of work has emerged from both the clinical and deep learning communities [4], [5], [6], [7] to directly evaluate AS from two-dimensional echo data. This enables evaluation to be accessible to a larger population in two ways: by easing the workflow of screening for AS and, more importantly, by allowing screening to be completed without spectral doppler.

Anatomical evaluation of the AV involves two standard-plane echo views, the parasternal long-axis (PLAX) and parasternal short-axis AV level (PSAX-Ao) (Figure 1), through which the AV is visible from two angles. These two views provide information on the structure of the valve, degree of calcification, speed and range of motion, all of which have an impact on the severity of AS. While apical views also provide visualization of the aortic valve, the opening of the aortic valve may not be clearly visible on the apical 5-chamber and apical 3-chamber views. A normal AV, as shown in Figure 1, does not show signs of thickening or calcification and fully opens, thus blood flows out of the heart without obstruction [8]. With the progression of AS, the AV thickens, its opening narrows, and its motion becomes more restricted. To automatically assess AS severity, a machine learning model should be able to focus on a few pixels in an echo image representing the AV, assess the AV's calcification and thickness, and understand the mobility of cusps throughout the cardiac cycle, all of which make this a fundamental and difficult task in video understanding.

Manuscript received 18 April 2023; revised 23 June 2023 and 4 August 2023; accepted 11 August 2023. Date of publication 15 August 2023; date of current version 2 January 2024. This work was supported in part by the Canadian Institute of Health Research (CIHR) and in part by the Natural Sciences and Engineering Research Council of Canada (NSERC). *(N. Ahmadi and M. Y. Tsang are co-first authors.) (Corresponding author: N. Ahmadi.)*

N. Ahmadi, A. N. Gu, and P. Abolmaesumi are with the Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver, BC V6T 1Z4, Canada (e-mail: nedaahmadi@ece.ubc.ca; guangnan@ece.ubc.ca; purang@ece.ubc.ca).

M. Y. Tsang and T. S. M. Tsang are with the Faculty of Medicine, The University of British Columbia, Vancouver, BC V6T 1Z4, Canada (e-mail: michael.tsang@ubc.ca; t.tsang@ubc.ca).

Digital Object Identifier 10.1109/TMI.2023.3305384

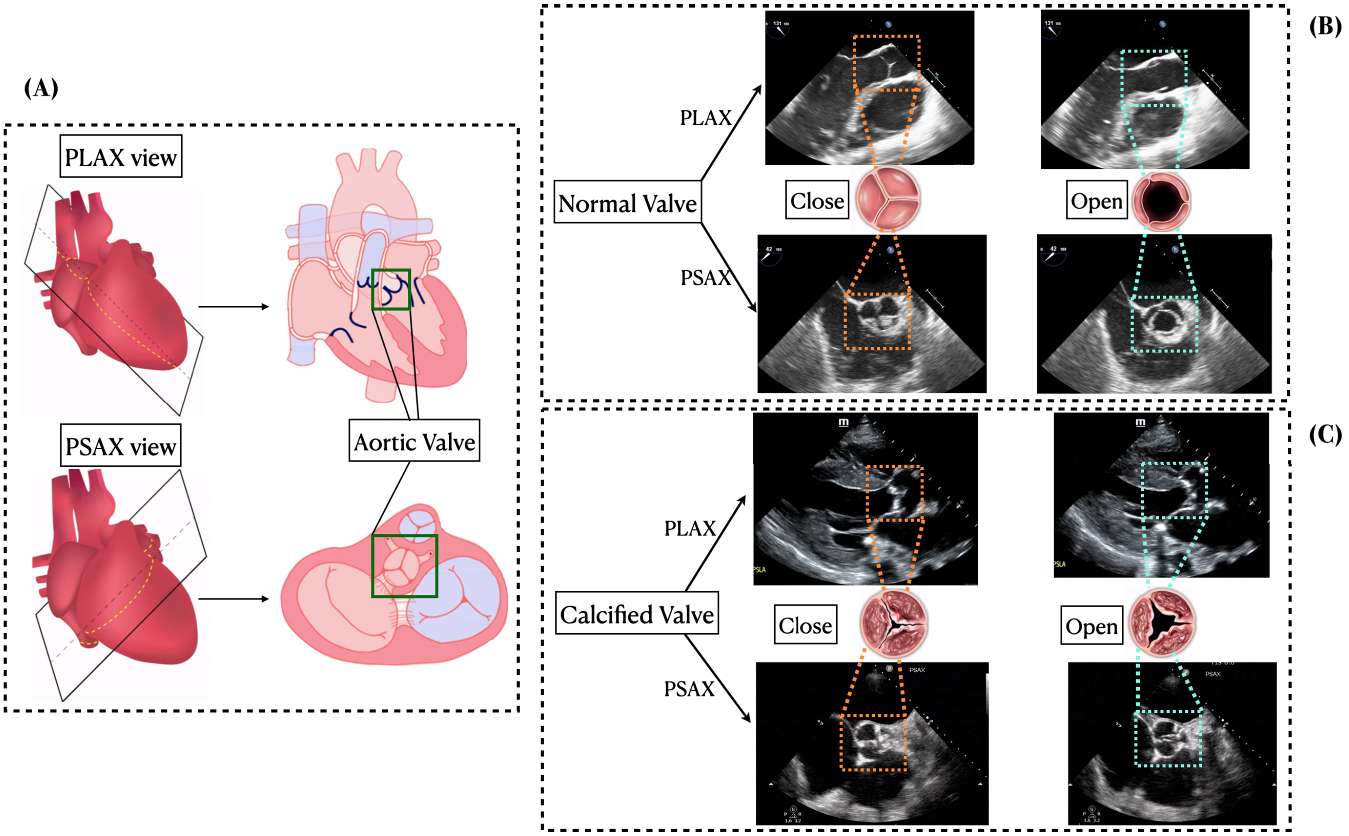

Fig. 1. (A) Diagram of the orientation of PLAX and PSAX views and their coincidence with the AV anatomy. (B) The appearance of the normal aortic valve in PLAX and PSAX views. The images on the left represent the closed AV, and the images on the right represent the open AV. (C) The appearance of calcified aortic valve in PLAX and PSAX views, and impact of calcification and narrowness of the valve on echo studies. Comparison of images in (B) and (C) demonstrates how calcification and thickening of the cusps present themselves in echo cine series, and how the progression of AS restricts the AV's motion.

Previous studies on automated AS assessment [5], [6] trained a deep neural network to learn the severity of AS from single echo images, then aggregated the predicted results of each image belonging to a patient using weighted averaging, where PLAX and PSAX views were assigned higher weights than other views. Based on our experiments and previous work [4], considering temporal information about valve opening and closing is also beneficial since the shape and mobility of the AV are the primary indicators of AS severity. We also observe that in most cardiac echo cine series, only a few frames show the opening and closing of the AV in an informative way that facilitates clinical decision-making. As a result, a simple video analysis model that is unable to pay attention to a subset of frames that are clinically relevant cannot provide an accurate classification of AS. Our problem is further complicated as each echo examination may contain multiple videos, which may not be equally informative of the AV structure or motion.

Based on the above observations, we investigated several approaches that leverage available literature on small object detection and temporal localization to tackle these challenges. Previous work has demonstrated video datasets provide additional temporal information that can be further incorporated for detecting small objects (e.g., [9], [10]) compared to methods that only consider spatial dimension [11]. The similarity of subsequent frames and slow changes in heart structure and background in echo enforces the need to capture local temporal context and small spatial changes of the aortic valve for a complete diagnosis. Additionally, to detect clinically informative frames in echo cine series, we took a look at temporal localization. Most current research [12], [13] are designed for action detection tasks, and use a weakly supervised learning method to identify the temporal interval of action classes. However, those methods are usually provided with a single or few frame-level annotations of whether a frame belongs to the background or is representing an action. When there is a lack of adequate temporal annotations (which is the case in clinical labels available for AS classification), several approaches have proposed unsupervised temporal localization for training action recognition networks [14], [15].

Inspired by those works, in this paper, we present a machine learning framework with the ultimate goal of developing Point-of-care EchocardioGraphy to detect AS with UltraSound (PEGASUS). Our framework has several key design features that facilitate training, including 1) using a temporal loss to enforce more sensitivity to small motions of AV in spatially similar frames without explicit AV localization labels; 2) adopting temporal attention to combine spatial representations with temporal context to capture the AV motion, which is reduced in the presence of moderate to severe stenotic valve;

and 3) automatically identifying the relevant echo frames that are more important for the final classification by learning from weak diagnosis labels and without explicit supervision. In summary, our contributions are as follows:

- We introduce an end-to-end spatio-temporal model with an efficient frame-level encoding that can learn small motions in echo by leveraging Temporal Deformable Attention (TDA) [16] in its transformer architecture. The model also adopts temporal coherence loss [17] to enforce detecting small spatial changes across frames.
- We introduce an attention layer to aggregate the disease severity likelihoods over a sequence of echo frames to produce a cine series-level prediction. These attention weights leverage temporal localization to find the most relevant frames in each cine series. We show that high attention weights consistently correlate with informative frames in each cine series.
- We demonstrate state-of-the-art accuracy on two clinical AS datasets, improving upon previous models for AS severity classification while having considerably less parameters compared to other video analysis models such as ResNet(2+1)D [18] and TimeSformer [19].

II. Related Work

The recent success of deep learning in analyzing medical imaging data, combined with the proliferation of medical imaging in clinical practice, are major motivators for the automation of AS diagnosis. This is particularly important in hospitals that are strained for staff or remote environments where access to cardiac imaging or expertise in cardiovascular medicine is sparse. These automated methods include the assessment of AS using a variety of data types.

A. Image Analysis

Kang et al. [20] used radiomics features from computed tomography AV calcium scoring (CT-AVC) to train a classifier for separating severe from non-severe AS, and noted the diagnostic accuracy is comparable to non-automated methods. Chang et al. [21] used deep learning to automatically segment calcified regions discovered by CT and predicted the severity of AS. Huang et al. [5] [6] applied a WideResNet [22] to predict the view and AS grading based on single two-dimensional echo images. They subsequently aggregated the predictions from each image belonging to the same patient to form a prediction at the patient level. Since most views are clinically uninformative and irrelevant, they conducted the final classification by a weighted sum of image-level logits, favoring the relevant views such as PLAX and PSAX.

B. Video Analysis

Roshanitabrizi et al. [23] used Doppler data of the PLAX and PSAX views to detect rheumatic heart disease (RHD), another pathology that can affect the AV. An ensemble method of 3D-Convolutional Neural Networks (CNN) and a transformer classify between normal and RHD cases. In point-of-care ultrasound devices, however, spectral Doppler is not generally available. Ginsberg et al. [4] proposed a video analysis approach to AS severity grading using two-dimensional echo cine series of the PLAX and PSAX views. They used a multi-task, uncertainty-aware training scheme with ResNet-18 2+1D [18] as the backbone model. They showed that multi-task training improves the model's generalization. This network cascades 1D temporal convolutions with 2D spatial convolutions. However, their work assumed each portion of the video is equally informative; thus, the impact of each frame on the final classification cannot be visualized or weighted accordingly. Dai et al. [24] uses 3D convolutional networks to estimate three Doppler measurements to detect AS severity levels. Vimalesvaran et al. [7] detected the presence of AS and aortic regurgitation using cardiac MRI cines. The algorithm is first trained on supervised key-point labels of the AV leaflets and blood flow jets, which are visible on MRI. An expert system and random forest performed feature extraction on the key-points and predicted pathology, respectively. Compared to fully deep architectures, their method is more interpretable.

In this work, we introduce a hand-crafted transformer-based architecture that is trained in an end-to-end approach and captures slight motions of AV without requiring any key-point labels while providing attention weights that represent the informativeness of frames within each cine.

III. Method

A. Model Overview

The overall architecture of our proposed framework is shown in Figure 2. Within every batch comprising B elements, given an input video (i.e. echo cine series) of arbitrary length, $X \in \mathbb{R}^{F\times H\times W\times 3}$, each frame is first encoded to a D-dimensional vector using a ResNet-18 based encoder. Frame-level feature vectors are concatenated to form a sequential representation $X_v \in \mathbb{R}^{F\times D}$. The features extracted from the video are then fed to a temporal encoder to capture the temporal context in the input feature sequence. In the final layer, the network is divided into three branches. The first branch calculates attention weights [14] using a fully connected layer, which provides an importance score for each frame. Class-specific confidence scores are derived from the second branch, which are then aggregated favoring attention weights with higher values to provide probability distribution for each video. The third branch provides a temporal loss, which ensures that the small local changes among subsequent frames are encoded in embeddings. Overall, the model is trained with a weighted sum of three losses:

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{cross\_entropy}\\ + \alpha\mathcal{L}_{attention\_entropy} + \beta\mathcal{L}_{temporal\_coherent}.\end{aligned} \tag{1}$$

All losses backpropagate through the same network, and hyperparameter values are identified based on the impact of each term on the total loss and refined using an empirical hyperparameter search.

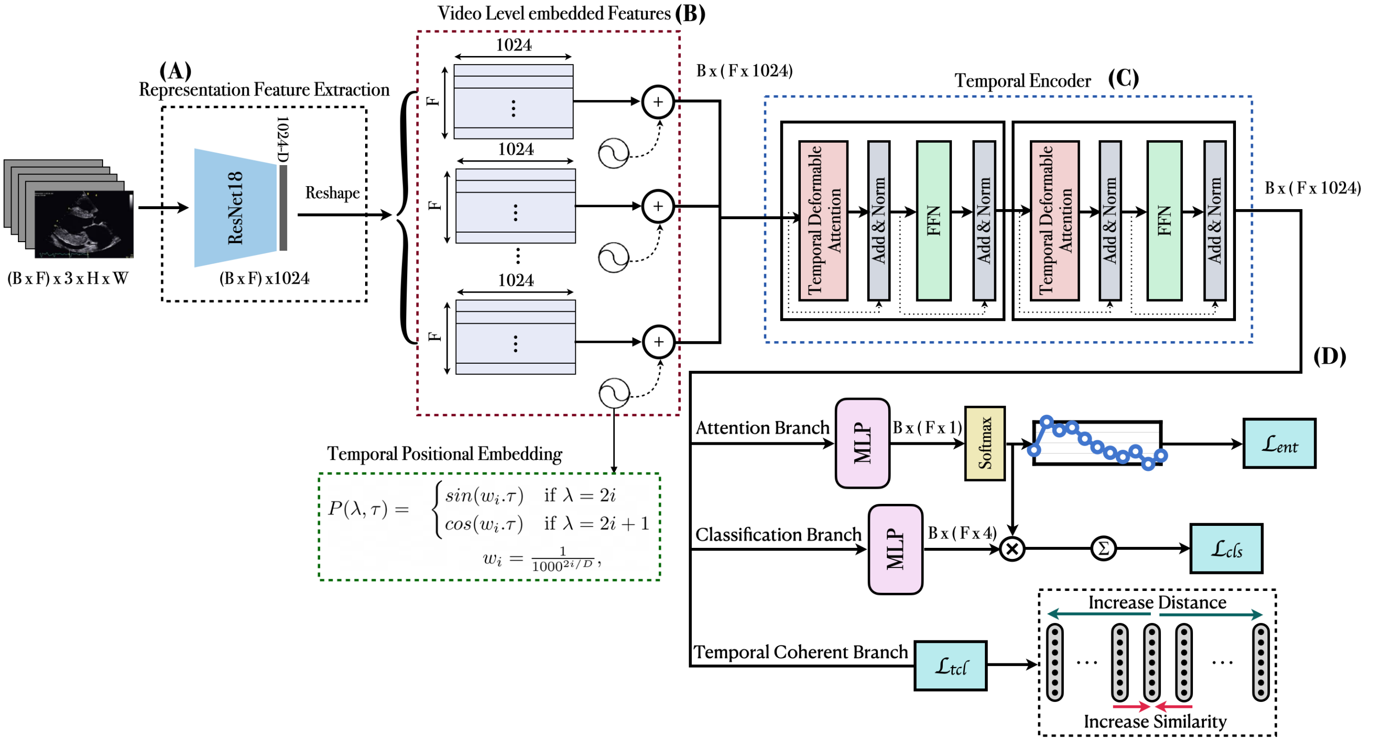

Fig. 2. Overview of the proposed machine learning framework. (A) Embeddings are extracted from each frame. (B) Extracted embeddings from frames of each ultrasound cine series are concatenated to create cine series level embedded features with the addition of temporal positional embedding. (C) The temporal encoder processes the temporal relation of embeddings. (D) Output embeddings are mapped to each class using attention weights. The total loss backpropagates into the whole network. In this context, B represents the number of elements in the batch, F represents the number of frames in the video, and H and W represent the height and width of each frame, respectively.

B. Temporal Positional Embeddings

The temporal position and order of frames are essential for accurate video understanding. In typical attention architectures, the attention module would perform identical inference on all frame-level embeddings, which does not provide information about the temporal relationships of the input frames.

Consequently, we use positional embedding based on time steps to provide order and temporal context to the input frames. We leverage from positional embedding used in [25] to encode this order in each video feature such that

$$P(\lambda,\tau) = \begin{cases} sin(w_i.\tau) & \text{if } \lambda = 2i \\ cos(w_i.\tau) & \text{if } \lambda = 2i+1 \end{cases}$$
$$w_i = \frac{1}{1000^{2i/D}}, \tag{2}$$

where τ = 1,…,F represents temporal position, and λ = 1,…,D represents location of each instance in an embedding. Thus, each time step in the sequence has a unique encoding, and the distance between two-time steps is consistent even for videos of different lengths. The process of incorporating positional information into the frame-level embeddings involves addition of variable P to the existing embeddings. This step results in an updated representation that reflects the temporal relationships of the frames within the sequence:

$$X_e = X_v + P. \tag{3}$$

C. Temporal Encoder

Much of the information related to AS severity is derived from the clinical assessment of echo videos, such as the opening and closing of the AV and the motion of the heart chambers. The temporal encoder uses temporal deformable attention (TDA) to enhance frame-level features with temporal information from nearby frames. Overall, the encoder consists of two transformer encoder layers inspired by [16], which replaces the dense attention found in typical transformer models with TDA followed by a feedforward network. Similar to the vanilla transformer architecture [25], outputs of each sublayer are fed to a residual connection and normalization layer.

1) Temporal Deformable Attention: Unlike action recognition tasks where an action can be seen in spatially distant frames, the AV motions observed in echo cine series are both small and local. To mitigate this issue, we take advantage of TDA (Figure 3). This attention module only samples small sets of key temporal locations around chosen reference points, independent of embedding size. Given an input video feature $X_e \in \mathbb{R}^{F\times D}$, for each query with index q and feature $v_q \in \mathbb{R}^D$ and its normalized position in time $t_q \in [0, 1]$, where 0 corresponds to the first frame and 1 corresponds to the last frame, the TDA feature is defined as

$$h_m = \sum_{k=1}^{K} a_{mqk}.W_m X_e((t_q + \Delta t_{mqk})F), \tag{4}$$
$$TDA(v_q, t_q, X_e) = W^o Concat[h_1, h_2, \ldots, h_m]_{m=1}^{M}, \tag{5}$$

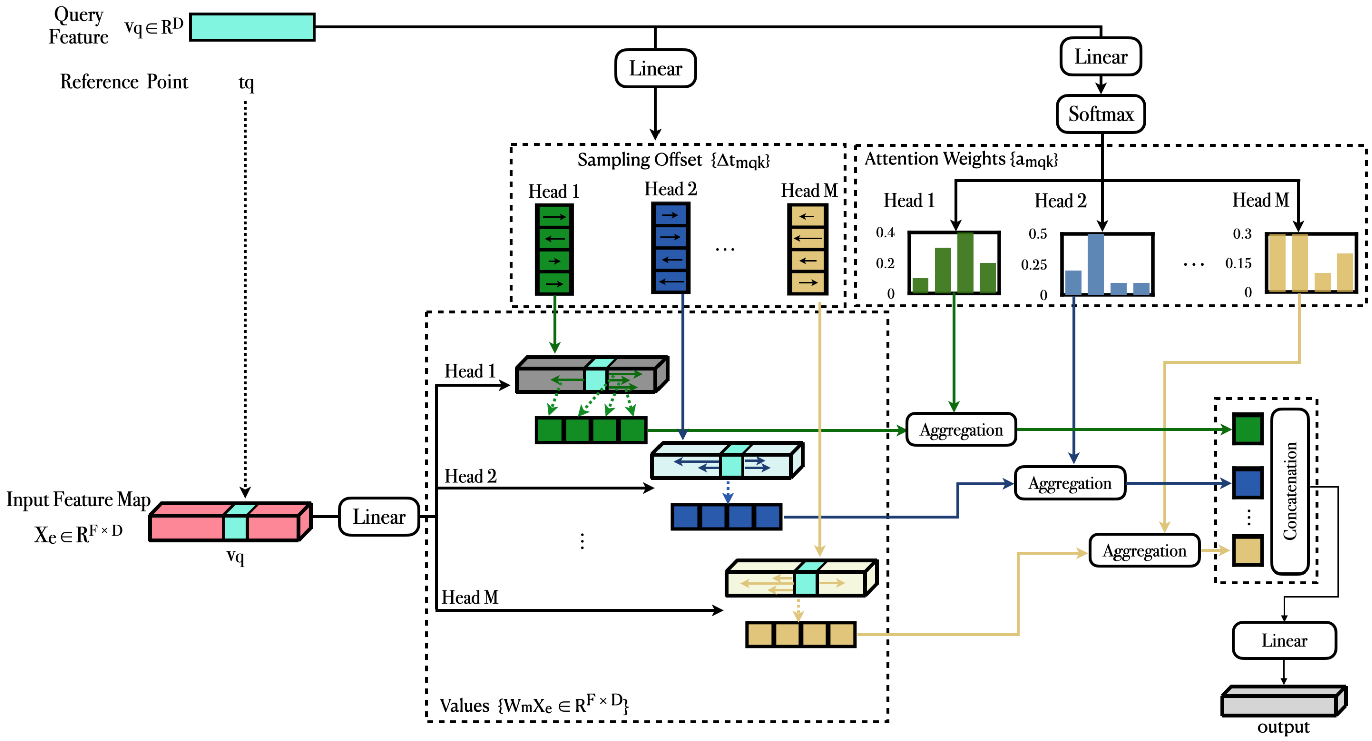

Fig. 3. Illustration of the temporal deformable attention (TDA) module. The input is an $F \times D$ matrix, where each vector represents one frame in the video. For each reference point in the temporal sequence, two linear projections are applied to the query feature $v_q \in \mathbb{R}^D$. The first branch encodes a small set of temporal offsets, which are then used to obtain key temporal locations. Normalized attention weights are derived by applying a softmax operator to the output of the second branch. The sampled key-points select elements from values which is a linear projection of input. Selected elements are then aggregated using attention weights for each attention head. The values are concatenated and fed into a linear projection to calculate the output. We show only one reference point and four sampled keys for a clear presentation.

where h_m is the output of the m-th head of TDA, a_{mqk} is the attention weight of the m^{th} sampling point in the k^{th} attention head for the q^{th} query. It is computed by performing a linear projection on each query, v_q, and subsequently normalizing the resulting values using a softmax function ($\sum_{k=1}^{K} a_{mqk} = 1$), $W_m \in \mathbb{R}^{D \times D/M}$ and $W^o \in \mathbb{R}^{D \times D}$ are the learned weights, m is the index of the attention head, M is the total number of attention heads, k is the index of the sampled key, K represents the total number of sampled keys, and F is the scalar video length. Δt_{mqk} is the sampling offset w.r.t t_q for the k^{th} sampled key and m^{th} attention head. To look up the value, we access X_e at the $(t_q + \Delta t_{mqk})F$-th position. Since $(t_q + \Delta t_{mqk})F$ may be a decimal, we use bi-linear interpolation in the time dimension on elements of X_e.

D. Attention Branch

In this branch, attention weights are calculated by applying a Multi Layer Perceptron (MLP) to output embeddings of the temporal encoder module. Attention weights are normalized via softmax along the temporal dimension. The weights indicate the importance of each frame in the final diagnosis probabilities. Since the frame-to-frame differences caused by AV motion can be small, the differences between frame embeddings are generally small too. In our design, we discourage attention weights from being too similar for each frame in the video. To achieve this, we add an entropy loss term based on the normalized weights to encourage sparsity:

$$\hat{\alpha} = \sigma_F(\alpha),$$
$$\mathcal{L}_{attention\_entropy} = -\sum_{\tau=1}^{F} \hat{\alpha}_\tau log(\hat{\alpha}_\tau). \quad (6)$$

where σ_F denotes softmax normalization across the temporal dimension F, $\alpha \in \mathbb{R}^F$ and $\hat{\alpha} \in \mathbb{R}^F$ are the attention weights before and after normalization, respectively.

E. Classification Branch

In order to derive the final cine series-level prediction, we use the attention weights for a weighted sum of class-specific logits. The probabilistic distribution of each class is calculated as follows:

$$p(y = c|x_{1:F}) \propto \sum_{\tau=1}^{F} \hat{\alpha}_\tau \sigma(f_{\theta_D}(x_\tau))_c, \quad (7)$$

where $f_{\theta_D}(.)$ is the output of the classification branch, and $\sigma(.)$ denotes softmax across classes.

For patient-level classification, we utilize entropy as an aleatoric uncertainty measure by employing four probabilities obtained from cine-level prediction. This allows us to assess the informativeness of each video. However, videos with an entropy value exceeding 0.3 are excluded from the analysis. Then we use majority voting to derive the final patient-level classification based on instances the model is more confident on. In cases where there is a tie, the maximum severity between those classes is selected.

F. Temporal Coherent Branch

Ideally, the frame features are consistent (i.e. have low variation) for adjacent frames but are still diverse as a distribution. To induce this property, we introduce a loss inspired by SyncNet [17], which tries to increase the similarity between adjacent frames and the distance between distant frames. This loss forces the model to create more distant embeddings for frames with small spatial differences such as those in our dataset. We formulate this loss as below:

$$
\begin{aligned}
L_{tcl} &= \frac{1}{F}\Big(\sum_{\tau=1}^{F} -log\big(\frac{e^{s_\tau}}{e^{s_\tau} + \sum_w e^{d_{\tau,w}}}\big)\Big) \\
s_\tau &= \begin{cases} v_{f_1}^T v_{f_2} & \text{if } \tau = 1 \\ v_{f_{F-1}}^T v_{f_F} & \text{if } \tau = F \\ \frac{1}{2} v_{f_{\tau-1}}^T v_{f_\tau} + \frac{1}{2} v_{f_\tau}^T v_{f_{\tau+1}} & \text{otherwise} \end{cases} \\
d_{\tau,w} &= v_{f_\tau}^T v_{f_w} \quad if \ |\tau - w| > T,
\end{aligned} \tag{8}
$$

where v_{f_τ} is the feature that represents frame f_τ, s_τ is calculated using the inner product of temporally adjacent frames, $d_{\tau,w}$ is the inner product of distant frames, w ranges from 1 to F, and T is the minimum temporal distance, measured in frames, that is considered distant. T was assigned three in our experiments. The computation of TCL is quadratic with respect to the number of frames due to the need to compute at least $F - 2T$ and at most $F - T$ similarities to find $\sum_w e^{d_{i,w}}$, but its has a low impact on runtime.

G. Dataset

We conduct experiments on two datasets: 1) a private video dataset, and 2) the TMED-2 [6] public image dataset, for AS classification and grading AS severity.

1) Private AS Dataset: The private dataset was sourced from a university-affiliated tertiary care hospital. Data were extracted with permission from the Information Privacy Office and the Clinical Medical Research Ethics Board. Cines were extracted from Philip IE33 and VividE9 ultrasound machines. In accordance with the American Heart Association Guidelines [3], AS severity levels were determined based on the three markers related to AS, namely AV area, peak valvular jet velocity, and mean pressure gradient provided in echo reports, resulting in an equal distribution of normal, mild, moderate, and severe cases. Furthermore, we only included studies with at least one PLAX or PSAX view and agreement between the calculated AV area and other Doppler parameters in terms of AS severity grading. In this proof-of-concept study, the exclusion of discordant cases refined our data and facilitated the development of a well-trained machine learning model. To establish the generalizability of this model, future studies will evaluate its performance in a wider population of individuals with aortic stenosis.

The two-dimensional PLAX and PSAX cine series of the selected studies were extracted from the hospital Picture Archiving and Communication System as follows. The echo data were anonymized in the hospital; all patient-identifying information and the echo-cardiogram tracing were removed from frames by applying a cine-shaped mask over the two-dimensional echo recording. We also removed any videos containing color or spectral Doppler. A deep-learning based view classification method [27] was used to automatically select only the PLAX and PSAX view videos. Finally, an experienced echocardiographer manually reviewed each video and removed videos from our dataset with the wrong view classification. The resultant dataset consists of only PLAX and PSAX videos and includes 2247 patients and 9117 videos.

To apply the data to our machine learning method, we divided the videos into training, validation, and test sets of approximately 70%, 20%, and 10%, respectively, ensuring mutually exclusive patients in each set. We extracted approximately one cardiac cycle from each video based on the patient's heart rate, and applied bilinear interpolation to resample the video to 32 frames. Subsequently, we resized each video to a spatial dimension of 224 × 224. We normalized the pixel intensities to zero mean and a standard deviation of 1. Finally, for the training set, we augmented the data using random horizontal flipping, rotation with the center on the beam origin and random cropping.

2) Public AS Dataset: TMED-2 dataset [6] consists of transthoracic echo studies from the Tufts Medical Center from 2011-2020. Each study contains multiple videos from various views, and studies are graded using Doppler-based guidelines [3]. Subsequently, they group the severity of AS into three categories: no AS, early AS (mild, mild-to-moderate), significant AS (moderate, severe). From each video, they extract the first frame as a representative image, and provide a label for the image view: PLAX, PSAX, 4-chamber, 2-chamber, and other. For each patient, around 50 to 100 images are available. The dataset contains three groups of images with respect to labels provided by board-certified sonographers or cardiologists, which are as follows:

- Fully-labeled set: Images from 577 patients for which both image-level view labels and patient-level AS severity are given;
- View-only labeled set: Images of 703 patients for which only view labels are given;
- Unlabeled set: 5287 patients without view or severity labels.

In this study we only used the fully-labeled set, DEV479, to compare to the baseline set by Huang et al. [6]. The train/test split was determined utilizing the generated csv file from the labeled dataset.

H. Implementation Details

Firstly, we use a ResNet-18 [28] backbone for representation feature extraction. We replace the final layer of the base model with a linear layer to yield feature vectors of dimension 1024. The feature maps of each video are stacked to form video features of size $F \times 1024$. Video features are fed to the temporal encoder. For the TDA sublayers, we use attention heads $M = 8$ and sampling points $K = 4$. The overall loss is

TABLE I

Test Accuracy Comparison With State-of-the-Art on Our Private AS Cine Series Dataset. Quantitative Results Show Our Approach Outperforms the State-of-the-Art in Both Video-Level and Patient-Level Classification. AS Severity Is a Four-Way Classification Encompassing the Classes of Normal, Mild, Moderate, and Severe, While AS Detection Entails a Two-Tier Classification Involving Normal Cases Versus All Other Severity Levels

			AS Severity Accuracy↑		Weighted F1-Score↑		AUROC↑	AS Detection Accuracy↑
Method	Model	Number of Parameters	Video-level	Patient-level	Video-level	Patient-level	Video-level	Patient-level
Imaged-based models								
Huang et al. [6]	WideResNet	23M	65.6%	66.70%	65.4%	66.60%	0.753	91.4%
Video-based models								
-	TinyVideoNet [26]	11M	65.7%	69.5%	65.5%	69.4%	0.754	93.2%
Ginsberg et al. [4]	ResNet(2+1)D	31M	68.4%	72.2%	67.4%	74.1%	0.777	94.2%
-	TimeSformer [19]	110M	68.8%	75.3%	67.9%	75.1%	0.779	94.7%
Ours	Ours	21.3M	**69.9%**	**78.1%**	**69.7%**	**78.0%**	**0.789**	**95.4%**

weighted with $\alpha = 0.01$ and $\beta = 0.1$. The model is trained using *Adam* [29] with an initial learning rate of 0.0001 and Cosine Annealing [30] as the learning rate schedule. For private dataset experiments, we train the model for 100 epochs. The model is developed using PyTorch [31] and experiments are conducted on two 16 GB Nvidia Titan GPUs. The hyperparameter optimization focused on the number of attention heads, keys within the transformer module and the weights used to aggregate loss functions. Furthermore, the metric used to guide the hyperparameter search was the accuracy of video-level AS severity.

I. Quantitative Results

Table I summarizes the test accuracy achieved by our method and various other state-of-the-art methods on the private dataset. We compare the accuracy of individual video classification and patient classification using multiple videos. See subsection III-E for the approach to combine predictions from multiple echo cine series. Our model outperforms other recent state-of-the-art methods while having smaller number of parameters compared to ResNet(2+1)D [18] and TimeSformer [19]. The accuracy and efficiency of our method suggest the effectiveness of explicitly considering the temporal dimension, especially with reference to the short-term nature of relevant AV motion, compared to vanilla video analysis architectures. Our findings demonstrate that AS detection accuracy is substantially higher than AS severity grading accuracy, largely due to normal cases being easier to classify and thin based on valve appearance, with blood flow obstruction differences. Conversely, diseased valves are usually calcified, and the extent of calcification and the constriction of valve motion vary, leading to differences in severity levels. As a result, differentiating between moderate AS and mild or severe cases is more difficult due to visual similarity. Furthermore, various factors such as noise, blurriness, or darkness of frames can obscure the aortic valve in many videos, making it challenging to assess its condition. Therefore, developing a model that can accurately and reliably classify most videos remains a difficult task.

J. Qualitative Results

1) Clinical Importance of Attention Weights: Our results support that learned attention weights have a direct correlation with temporal clinical information. This is shown in Figure 4. Most frames that represent open AV have higher weights, and the lowest weights are associated with closed AV. We hypothesize that the network is taking advantage of the valve motion and changes in its shape during the cardiac cycle to make its prediction. Following the addition of the attention entropy loss, the model exhibited a greater degree of attention sparsity, indicating a more focused allocation of attention across the frames.

2) Coherency of Embeddings: We also analyzed the learned features from videos by demonstrating the impact of temporal coherent loss on the similarity and distinctness of the frames in each cine series. Figure 5 illustrates that embeddings that belong to the same stage of a heart cycle are more similar to each other and more distant to frames that represent another phase of the cycle.

3) Failure Study: We applied an extensive failure study on patients with a large number of mislabelled videos. Videos are often misclassified with the presence of noise, the darkness of frames, poor image quality, and invisibility of the AV and its cusps. We visualized three failure cases from the test set of our private dataset in Figure 6. For all three samples, there is a two-level difference between the prediction and ground truth, which can largely impact clinical outcomes. In the first failure example, the AV is visible in all frames; however, the cusps cannot be seen. Therefore, the calcification and the narrowness of open valve cannot be estimated accurately. But, as the layout of the valve is clear, the attention weights have assigned higher weights to frames with an open AV. In the second example, most frames are dark, so, they provide little clinical information. As a result, large sections of the heart structure and small motions are undetectable. This similarity among frames and its difference with common cine series in the training set resulted in fairly similar and uninformative embedding space. The third example shows a video of good quality in which the AV is not visible. Again, attention weights were able to detect frames representing heart contraction.

K. Empirical Ablation Analysis

The contribution of each model component was analyzed by performing an ablation study. Different components were eliminated or replaced; video-level and patient-level accuracy were used to compare different settings (see Table II).

1) Impact of Each Layer on Representation Extraction: Replacing the ResNet-18 encoder with ResNetAE based on the

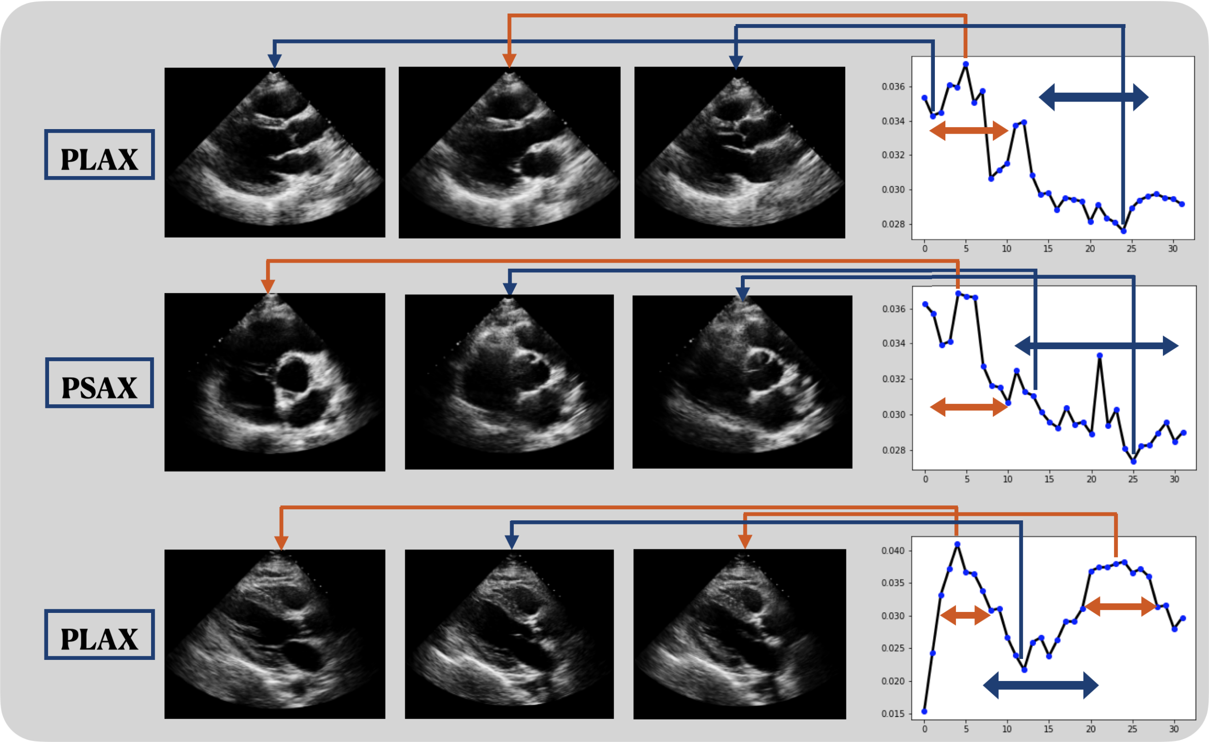

Fig. 4. Qualitative examples of how attention weights have learned the informativeness of frames. The diagram on the right shows the attention weights associated with 32 frames of each video. Three sample frames of each video are shown on the left side. The orange arrows show the interval of frames with an open AV and its associated attention weights. Blue arrows represent the close AV both before and after heart contraction.

work of [32] showed that the ResNetAE embeddings provide a good representation of each frame, but the temporally distant frames in each video produced similar embeddings due to the similarity of their spatial dimension. This prohibited the model from learning the temporal variation throughout the cine. We experimented with both temporal transformer encoder (TTE) and BERT architectures for temporal encoding. We observed that BERT could not capture the small local changes between the embeddings of adjacent frames. We validated this by comparing the accuracy between a model with ResNet-18 and the BERT encoder and a model using only the ResNet-18 layer without the transformer encoder that averaged the embedding for all frames. We observed that the change in accuracy was not significant when we added the BERT encoder. This indicates that the BERT encoder was unable to capture the temporal information. However, using the temporal transformer encoder resulted in a 4.6% increase in video-level accuracy. This indicates the notable impact of replacing dense attention with TDA.

2) Aggregation Method: We tested two aggregation methods to calculate the class-level probabilities. In the first method, all logits are averaged, which disregards the importance of each frame in the cine series and its impact on the final diagnosis. In the second method, we used normalized attention weights as a weighted score to combine class-specific predictions. Our experiments show attention weighting, even without the attention entropy loss, yields slightly better accuracy.

3) Pretraining Weights of Encoder: We tried pre-training weights of ResNet-18 using supervised contrastive loss (SupCon) to learn more informative representations. Then, we froze its weights during training. However, this did not result in any improvement. Based on our experiments, we conclude that good representation extraction is not sufficient, and our empirical studies validate the advantage of end-to-end learning, especially with the impact of temporal coherent loss on learning better overall representations.

4) Impact of Each Loss Function: As we can see in Table II, the attention entropy (AttE) loss only improves the accuracy by 0.4%. However, before adding the loss, weights assigned to each frame were more similar. Therefore, this loss has a positive influence on the sparsity of informative frames. Since the frames of a cine are visually similar due to small changes caused by muscle contraction and valve movement, for certain examples, we have observed that there was a lack of significant differences in attention weights. However, in most samples, frames that show an open aortic valve have higher weights assigned to them after addition of the loss. The temporal coherent loss (TCL) improves the accuracy by 0.7%. Finally, we use all losses to train the action localization model, and achieve an accuracy of 69.4%, implying that each loss contributes to the overall accuracy.

L. Evaluation of Our Method on Public Dataset

1) Reproducing the Baseline: Due to the lack of a large public video dataset with AS diagnosis labels, we tested

TABLE II

ABLATION STUDY OF NETWORK COMPONENTS ON THE VALIDATION SET OF OUR PRIVATE DATASET, STUDYING THE IMPACT OF EACH COMPONENT. CE: CROSS ENTROPY. ATTE: ATTENTION ENTROPY. TCL: TEMPORAL COHERENT LOSS. SUPCON: SUPERVISED CONTRASTIVE LEARNING

Model Architecture				Loss Functions			Accuracy↑	
Image Encoder	Transformer	Weakly-supervised Aggregation	Pretraining	CE	AttE	TCL	Video Level	Patient Level
ResNetAE	BERT	Averaging	-	✓	×	×	61.9%	64.5%
ResNet-18	BERT	Averaging	-	✓	×	×	63.6%	65.9%
ResNet-18	BERT	Averaging	SupCon	✓	×	×	65.8%	67.3%
ResNet-18	TTE	Averaging	-	✓	×	×	68.2%	74.2%
ResNet-18	TTE	Averaging	SupCon	✓	×	×	67.8%	73.0%
ResNet-18	TTE	Attention Weight	-	✓	×	×	**68.6%**	**75.1%**
ResNet-18	TTE	Attention Weight	-	✓	✓	×	68.7%	75.3%
ResNet-18	TTE	Attention Weight	-	✓	×	✓	69.0%	76.7%
ResNet-18	TTE	Attention Weight	-	✓	✓	✓	**69.4%**	**77.8%**

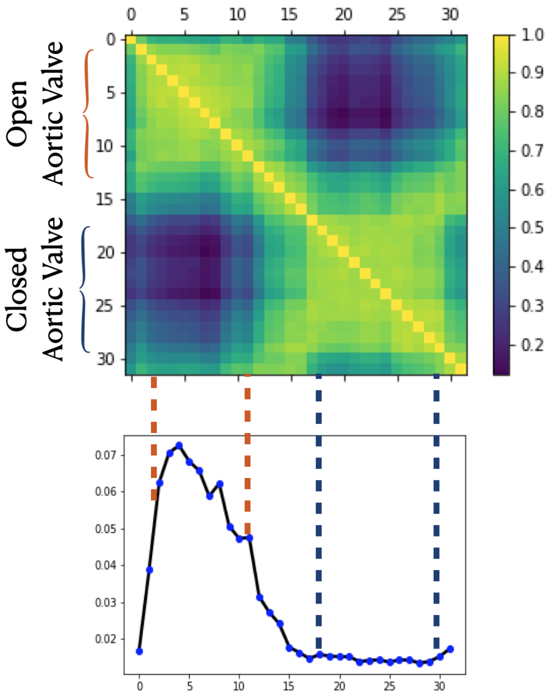

Fig. 5. The upper figure illustrates the pairwise similarity of frame-level representations based on their cosine distance. The lower figure exhibits the attention weights of each frame. As shown, similar frames have been divided into two subgroups. The first group represents a phase of the heart cycle with an open AV. We can see that these frames also have higher attention weights. Comparatively, the second subgroup mostly belongs to frames with closed AV and they have lower attention weights.

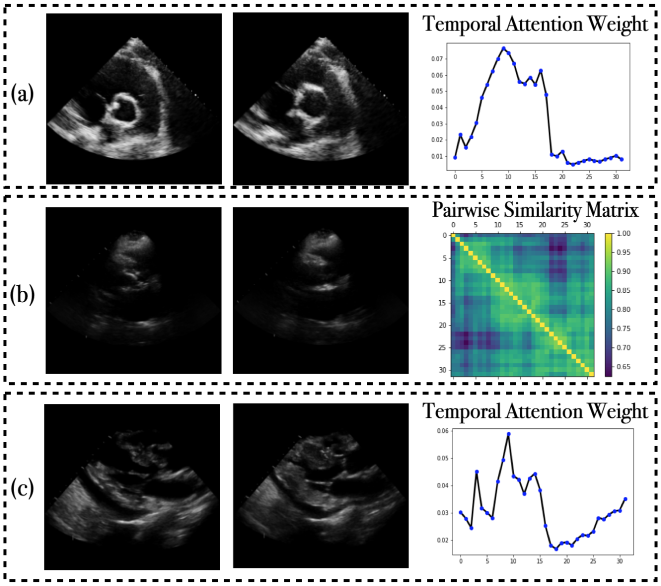
(a)
Temporal Attention Weight
(b)
Pairwise Similarity Matrix
(c)
Temporal Attention Weight

Fig. 6. This figure shows three failure cases, with representative frames of a closed and an open AV in order from left to right. (a) Visible AV with undetectable cusps due to noise. However, attention weights were able to detect phases of the heart cycle but not calcification. (b) Fairly dark and uninterpretable frames. Fairly similar embedding as a result. (c) Good video quality, but in most frames AV can not be detected. Still, because of having fairly good quality video, attention weights could detect frames with open AV.

our attention aggregation method on the TMED-2 [6] image dataset. To reproduce their results, we resized each image to 224 × 224. We trained a multitask WideResNet-50-2 network [22] to provide a label for image view (PLAX, PSAX, and others) and severity of AS (no AS, early AS, i.e. mild or mild-to-moderate, and significant AS, i.e. moderate or severe). At inference time, images with high view classification entropy were disregarded for each patient, and the summation of the probability of relevant views (PLAX and PSAX) was calculated, where thresholding was used to select images with a high likelihood of belonging to one of the clinically relevant views. The weights of the selected images were adopted to perform a weighted aggregation:

$$r\_w(x) = \sigma(f_{\theta_v}(x))_{PLAX} + \sigma(f_{\theta_V}(x))_{PSAX}$$

$$\hat{r}\_w(x) = \begin{cases} 0 & \text{if } r\_w(x) < \tau_1 \\ 0 & \text{if } H(r\_w(x)) > \tau_2 \\ r\_w(x) & \text{else} \end{cases}$$

$$p(y = c|x_{1:n}) \propto \sum_{i=1}^{n} \hat{r}\_w(x_i)\sigma(f_{\theta_D}(x_i))_c, \qquad (9)$$

where f is the network, θ_v is the view classifier parameters, θ_D is the AS diagnosis classifier parameters, $\sigma(.)$ denotes softmax, r_w and $\hat{r_w}$ are the relevance weight before and after thresholding, τ_1 are the confidence thresholds of belonging to relevant views, and τ_2 is a threshold for having a low entropy for the predicted probabilities. With our implementation, we were able to obtain slightly better results compared to those reported

TABLE III
PATIENT-LEVEL AS SEVERITY DIAGNOSIS CLASSIFICATION IN THE TMED-2 DATASET. COMPARISON WITH THE STATE-OF-THE-ART METHOD [6] AND THE DIFFERENCE IN AGGREGATING IMAGES FOR PATIENT-LEVEL DIAGNOSIS

Method	Aggregation	Patient-level Accuracy↑
Huang et al. [6]	thresholding, view relevance	74.6%
Reproduced Results	thresholding, view relevance	75.6%
Ours	attention based, view relevance	**83.8%**

in [6]. The values selected for τ_1 and τ_2 were 0.7 and 0.3, respectively.

2) Implementation of Our Method: As TMED-2 is image-based, we trained our model without considering the transformer layer and temporal coherent loss and replaced ResNet-18 with WideResNet to be able to compare the results. Since the attention module is trained on groups of images and the number of images per patient is variable, for each patient, these images were fed into the model for feature extraction. Three MLPs were applied on image-level embeddings to obtain attention weights, view classification, and AS classification. Since the attention module operates on groups of images belonging to the same patient, the network was trained at a patient-level where the number of images per patient is variable. To accommodate for variable-length input, we binned the patients based on the number of images and defined multiple data loaders, each for a different bin. Attention aggregation was used to obtain the severity of AS from the multiplication of the AS classification branch and view relevance. We also added the entropy loss for attention weights to learn more informative images. Patient-level training increased the accuracy of AS detection and AS severity classification to 91.5% and 83.8%, respectively. Compared to the aggregation of the image-level model at inference time, (see Table III), the addition of attention weights had a significant impact on the calculated probability distribution. One reason behind this may be that although only PLAX and PSAX views are clinically relevant, not all PLAX and PSAX images provide sufficient information to diagnose AS. The attention map can learn to choose more informative images during training:

$$p(y = c|x_{1:n}) \propto \sum_{i=1}^{n} \hat{\alpha}_i \hat{r}\_w(x_i)\sigma(f_{\theta_D}(x_i))_c, \qquad (10)$$

where $\hat{\alpha}$ are the attention weights normalized across images.

IV. CONCLUSION

In this work, we introduce a novel architecture for detecting the severity of AS in cardiac echo cine series. We demonstrate three architectural choices that resulted in more accurate detection and grading by: 1) leveraging from temporal deformable attention to increase locality awareness in transformers; 2) using temporal coherent loss to capture small spatial changes and enforce coherency in frame-level embeddings, and 3) adopting attention weights for detecting frames that provide clinical relevance and favoring those frames in weighted aggregation. We analyze the importance of each component in improving accuracy and outperforming state-of-the-art methods. For future work, we plan to extend this framework to find informative videos for patient-level classification. This may include leveraging uncertainty to disregard videos with insufficient clinical information. We aim to include interpretability as part of our design and to facilitate the adoption of the approach toward point-of-care ultrasound settings.

ACKNOWLEDGMENT

T. S. M. Tsang and P. Abolmaesumi are joint senior authors.

REFERENCES

[1] B. A. Carabello, “Introduction to aortic stenosis,” *Circulat. Res.*, vol. 113, no. 2, pp. 179–185, Jul. 2013.

[2] G. Strange et al., “Poor long-term survival in patients with moderate aortic stenosis,” *J. Amer. College Cardiol.*, vol. 74, no. 15, pp. 1851–1863, Oct. 2019.

[3] C. M. Otto et al., “Guideline for the management of patients with valvular heart disease: A report of the American heart association joint committee on clinical practice guidelines,” *Amer. College Cardiol. Found.*, vol. 77, pp. e25–e197, Jan. 2021.

[4] T. Ginsberg et al., “Deep video networks for automatic assessment of aortic stenosis in echocardiography,” in *Simplifying Medical Ultrasound*, 2021, pp. 202–210.

[5] Z. Huang, G. Long, B. Wessler, and M. C. Hughes, “A new semi-supervised learning benchmark for classifying view and diagnosing aortic stenosis from echocardiograms,” in *Proc. 6th Mach. Learn. Healthcare Conf.*, 2021, pp. 1–11.

[6] Z. Huang, G. Long, B. Wessler, and M. C. Hughes, “TMED2: A dataset for semi-supervised classification of echocardiograms,” in *Proc. Int. Conf. Mach. Learn. DataPerf Workshop*, 2022.

[7] K. Vimalesvaran et al., “Detecting aortic valve pathology from the 3-chamber cine cardiac MRI view,” in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2022*, 2022, pp. 571–580.

[8] L. Ring et al., “Echocardiographic assessment of aortic stenosis: A practical guideline from the British society of echocardiography,” *Echo Res. Pract.*, vol. 8, no. 1, pp. G19–G59, Mar. 2021.

[9] F. Xiao and Y. J. Lee, “Video object detection with an aligned spatial–temporal memory,” in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 494–510.

[10] L. Chen, W. Shi, and D. Deng, “Improved YOLOv3 based on attention mechanism for fast and accurate ship detection in optical remote sensing images,” *Remote Sens.*, vol. 13, pp. 2072–4292, 2021.

[11] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, “Deformable DETR: Deformable transformers for end-to-end object detection,” in *Proc. Int. Conf. Learn. Represent.*, 2021.

[12] P. Lee and H. Byun, “Learning action completeness from points for weakly-supervised temporal action localization,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 13628–13637.

[13] J. Ma, S. K. Gorti, M. Volkovs, and G. Yu, “Weakly supervised action selection learning in video,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7583–7592.

[14] G. Gong, X. Wang, Y. Mu, and Q. Tian, “Learning temporal co-attention models for unsupervised video action localization,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9816–9825.

[15] K. Soomro and M. Shah, “Unsupervised action discovery and localization in videos,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 696–705.

[16] X. Liu et al., “End-to-end temporal action detection with transformer,” *IEEE Trans. Image Process.*, vol. 31, pp. 5427–5441, 2022.

[17] F. T. Dezaki et al., “Echo-SyncNet: Self-supervised cardiac view synchronization in echocardiography,” *IEEE Trans. Med. Imag.*, vol. 40, no. 8, pp. 2092–2104, Aug. 2021.

[18] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, “A closer look at spatiotemporal convolutions for action recognition,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6450–6459.

[19] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" in *Proc. ICML*, Jul. 2021, pp. 813–824.

[20] N. G. Kang, Y. J. Suh, K. Han, Y. J. Kim, and B. W. Choi, "Performance of prediction models for diagnosing severe aortic stenosis based on aortic valve calcium on cardiac computed tomography: Incorporation of radiomics and machine learning," *Korean J. Radiol.*, vol. 22, no. 3, p. 334, 2021.

[21] S. Chang et al., "Development of a deep learning-based algorithm for the automatic detection and quantification of aortic valve calcium," *Eur. J. Radiol.*, vol. 137, Apr. 2021, Art. no. 109582.

[22] S. Zagoruyko and N. Komodakis, "Wide residual networks," in *Proc. Brit. Mach. Vis. Conf.*, 2016, p. 87.

[23] P. Roshanitabrizi et al., "Ensembled prediction of rheumatic heart disease from ungated Doppler echocardiography in low-resource settings," in *Proc. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2022, pp. 602–612.

[24] W. Dai, H. Nazzari, M. Namasivayam, and J. Hung, "Identifying aortic stenosis with a single parasternal long-axis video using deep learning," *J. Amer. Soc. Echocardiography*, vol. 36, no. 1, pp. 1–12, Jan. 2023.

[25] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–13.

[26] A. Piergiovanni, A. Angelova, and M. Ryoo, "Tiny video networks: Architecture search for efficient video models," Tech. Rep., 2020.

[27] A. N. Gu et al., "Efficient echocardiogram view classification with sampling-free uncertainty estimation," in *Simplifying Medical Ultrasound*, 2021, pp. 139–148.

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

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[30] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," in *Proc. Int. Conf. Learn. Represent.*, 2017.

[31] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 8024–8035.

[32] H. Reynaud, A. Vlontzos, B. Hou, A. Beqiri, P. Leeson, and B. Kainz, "Ultrasound video transformers for cardiac ejection fraction estimation," in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2021*, 2021, pp. 495–505.