



Spatiotemporal feature disentanglement for quality surveillance of left ventricular echocardiographic video using ST-R(2 + 1)D-ConvNeXt

Chin-Chieh Hsu^a, You-Wei Wang^a, Lung-Chun Lin^{b,*}, Ruey-Feng Chang^{a,c,d,*}

^a Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan

^b Section of Cardiology, Department of Internal Medicine, National Taiwan University Hospital, Taipei, Taiwan

^c Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taipei, Taiwan

^d Graduate Institute of Network and Multimedia, National Taiwan University, Taipei, Taiwan

ARTICLE INFO

Keywords:

Deep learning

Left ventricular echocardiography

Quality surveillance

Spatiotemporal feature disentanglement

ABSTRACT

The left ventricle (LV), as the primary chamber responsible for systemic circulation, plays a crucial role in cardiac function assessment. Echocardiography which particularly focuses on LV, is vital for cardiac disease diagnosis. However, the diagnostic accuracy heavily depends on image quality, which requires systematic assessment. In this study, we propose a two-stage deep learning approach for echocardiographic quality surveillance using a dataset of 514 annotated videos. The first stage employs EchoNet, to extract LV volumes of interest. The second stage introduces ST-R(2 + 1)D-ConvNeXt, a novel ConvNeXt-based model designed to disentangle spatiotemporal features and leverage echocardiographic hallmarks within the apical-four-chamber (A4C) dynamic echocardiogram data. The proposed approach achieves an accuracy of 82.63 %, an Area Under the Curve (AUC) of 0.89, a sensitivity of 84.10 %, and a specificity of 81.08 % in classifying echocardiographic videos into high and low quality. Furthermore, through explainable AI techniques, our model identifies specific quality issues such as missing cardiac walls, distorted or poorly positioned chambers, and other anomalies, providing interpretable feedback for clinical applications.

1. Introduction

The heart, a critical organ responsible for blood circulation and nutrient distribution, comprises four chambers: left ventricle (LV), right ventricle (RV), left atrium (LA), and right atrium (RA), as illustrated in Fig. 1. Among these chambers, the LV plays a crucial role in cardiac function assessment, as LV directly influences systemic circulation and is closely associated with various cardiovascular diseases [1].

Echocardiography, also known as ultrasound cardiography (UCG), has become the primary tool for cardiac function assessment due to its non-invasive nature, high portability, and cost-effectiveness. However, the diagnostic accuracy of echocardiography heavily depends on image quality, which is frequently compromised by factors such as acoustic shadows, speckle noise, and motion artifacts. Several studies simply remove low-quality images from the analysis [2,3], which may significantly limit the available data for studies and compromise the reliability of the analysis.

Recent advances have introduced automated analysis through view classification and image segmentation [4,5]. These approaches

primarily focus on static image features and largely overlook two critical aspects. First, they fail to utilize temporal dynamics of the heartbeat, which contain valuable information about cardiac wall motion and contractility. Second, they do not leverage the inherent characteristics of echocardiographic imaging, such as its monochromatic nature and topological consistency, which we term *echocardiographic hallmarks*. These hallmarks could potentially enhance the reliability of automated analysis.

The apical-four-chamber (A4C) view demonstrated a broader range of classification confidence values compared to A2C and A3C views, indicating superior discriminative capability for image quality assessment [6]. However, maintaining consistent image quality in A4C acquisitions remains challenging, often requiring multiple attempts and extensive operator expertise. These challenges, coupled with the increasing demand for efficient cardiac diagnostics, underscore the need for automated quality assessment systems specifically designed for dynamic A4C echocardiograms. The clinical significance of such automation is multifaceted. First, the automation serves as an early warning system for potentially unreliable measurements, enabling immediate

* Corresponding authors.

E-mail addresses: anniejou@ms28.hinet.net (L.-C. Lin), rfchang@csie.ntu.edu.tw (R.-F. Chang).

<https://doi.org/10.1016/j.bspc.2025.107671>

Received 5 November 2024; Received in revised form 24 January 2025; Accepted 7 February 2025

Available online 13 February 2025

1746-8094/© 2025 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

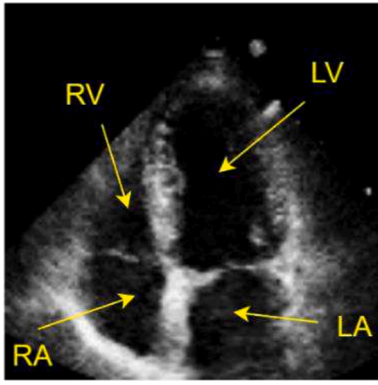


Fig. 1. Four-chamber view echocardiogram showing the four primary chambers of the heart: left ventricle (LV), right ventricle (RV), left atrium (LA), and right atrium (RA). Note that the image is flipped both horizontally and vertically compared to the human body in vivo due to the imaging angle.

image re-acquisition rather than discovering quality issues during subsequent diagnosis. Second, it provides objective quality metrics that can standardize image acquisition across different operators and institutions. Third, by identifying high-quality images with consistent manual and automated measurements, the diagnostic decisions by clinicians can be enhanced. This systematic approach not only improves clinical workflow efficiency but also reduces the need for repeat examinations, particularly crucial in time-sensitive clinical settings where rapid yet reliable assessment is essential.

Our work addresses these challenges through three primary contributions:

- 1) *We are the first to clearly separate features in dynamic echocardiography into two distinctive representations: spatiotemporal attributes and echocardiographic hallmarks, where the spatiotemporal attributes are further disentangled into spatial and temporal features in our model. We utilize them accordingly with specific justifications and achieve an accuracy of 82.63 % and an Area Under the Curve (AUC) of 0.89 in the quality classification task on the echocardiograms.*
- 2) *We develop an interpretable quality assessment system that not only identifies suboptimal images but also localizes and characterizes specific quality issues. This capability offers a fresh perspective on the explainable AI technique and directly supports clinical decision-making by providing actionable feedback for image acquisition optimization.*
- 3) *We delve into the area of echocardiography quality assessment, a potentially underexplored domain, and demonstrate the effectiveness of focusing on LV chambers through a two-stage assessment framework, establishing a new approach to echocardiographic quality surveillance that aligns with clinical workflow requirements.*

These contributions collectively advance the field of automated echocardiographic quality assessment, addressing a critical gap in current clinical practice while providing a foundation for more reliable cardiac diagnosis.

2. Related work

To establish a comprehensive foundation for echocardiographic quality assessment, we survey three critical aspects: 1) fundamental developments and challenges in quality surveillance, 2) the role of dynamic feature analysis in video processing, and 3) contemporary deep learning approaches in cardiac video analysis. We begin by examining quality surveillance methods, followed by an exploration of dynamic feature characteristics, and conclude with a review of modern architectural innovations and their potential challenges in echocardiographic video analysis.

2.1. Echocardiographic quality surveillance

Quality surveillance for medical images plays a crucial role in ensuring diagnostic accuracy, yet traditionally relies heavily on operator expertise [7], which inherently introduces varying subjective interpretations and could potentially compromise assessment consistency [8,9]. In fact, how various medical imaging technologies present fundamentally different technical challenges in quality evaluation has been highlighted: Magnetic Resonance Imaging (MRI) quality might be compromised by patient movement and field inhomogeneity, Computed Tomography (CT) faces distinct challenges from radiation-induced effects, and ultrasound contends with characteristic speckle patterns. These modality-specific variations in image degradation mechanisms and artifact patterns underscore the need for modality-specialized assessment approaches [10]. In the realm of echocardiography, representation learning becomes challenging due to the dynamic nature of cardiac motion and various imaging artifacts [11]. These complexities have underscored the necessity for automated quality assessment approaches, leading to increased research efforts in recent years.

Pioneering work introduced a Convolutional Neural Network (CNN) framework for real-time quality assessment of echocardiograms, demonstrating potential generalizability across different cardiac views [12]. This was further extended to incorporate 3D information through recurrent neural networks, marking an important step towards addressing the dynamic nature of cardiac imaging, which is necessary for image feature exploitation [7]. Building upon this foundation, Luong et al. developed and validated an automated quality estimation system precisely for hospitalized patients, addressing practical clinical needs [13].

However, recent findings suggest that direct 3D approaches for capturing dynamic representations in echocardiography may be sub-optimal. Specifically, Jansen et al. demonstrated that a 3D CNN architecture underperformed in the echocardiographic quality assessment task compared to a simpler 2D frame-by-frame CNN coupled with temporal averaging [14]. Despite the intuitive appeal of 3D architectures for handling dynamic cardiac sequences, their study revealed that the conventional frame-by-frame processing followed by temporal integration could provide both superior performance and computational efficiency. This property emphasizes the significance of temporal features over spatial representations, providing theoretical support for the effectiveness of sequential 2D processing approaches.

Quality evaluation methods, another important topic, have also undergone significant evolution. Left Ventricular Ejection Fraction (LVEF) and Global Longitudinal Strain (GLS) are the two most common metrics associated with LV [15,16]. However, numerous studies have indicated that LVEF may be limited due to lower correlation with cardiac diseases, such as heart failure [17], and GLS is preferred due to its notable reproducibility and robustness [18] and can better reflect human expert assessment of ultrasound image quality [19]. Building on this progress toward more clinically relevant evaluation, Huang et al. investigated the relationship between image quality and GLS measurements, showing that poor image quality significantly impacts the reliability of GLS calculations [6].

2.2. Dynamic feature analysis in echocardiography

The dynamic nature of echocardiography is a key characteristic that makes it well-suited for cardiac function assessment [20] and necessitates careful consideration of both spatial and temporal features. Painchaud et al. revealed that conventional CNN approaches often fail to adequately capture temporal information in cardiac sequences, emphasizing the importance of maintaining temporal consistency throughout the cardiac cycle, particularly for accurate segmentation and quality assessment [21]. Valuable insights have also been obtained from other cardiac imaging modalities.

In fact, the significance of disentangled representation learning is not

limited to echocardiography; in Cardiac Magnetic Resonance Imaging (CMRI) such concepts have also been established. The findings on separating anatomical and motion-related features have important implications for echocardiographic analysis [22], where similar principles can be applied to isolate quality-relevant features in dynamic cardiac images. This concept of feature disentanglement is particularly valuable in handling the unique characteristics of echocardiographic data. Unlike other medical imaging modalities such as MRI, echocardiographic data not only suffers from low image quality but also faces the challenge that different cardiac formations share similar spatial patterns, making the independent exploitation of temporal features particularly crucial [23]. This notion was pioneered by Gao et al., who proposed a two-stream CNN architecture that separately processes spatial and temporal features for viewpoint classification of echocardiograms, once again displaying that the independent analysis of temporal dynamics can outstandingly enhance the classification performance [24].

Consequently, recent studies have increasingly recognized the importance of temporal coherence and resolution in echocardiographic analysis [25,26]. Østvik et al. highlighted how incorporating temporal information improves myocardial function assessment [27]. More recently, Yang et al. exhibited the effectiveness of combining temporal motion artifact removal with spatial feature disentanglement in improving diagnostic accuracy of cardiac conditions, achieving performance comparable to that of experienced clinicians [28]. This significance was further validated by Mokhtari et al., who proposed a multi-level transformer framework that explicitly separates spatial and temporal features for different diagnostic tasks while maintaining clinical interpretability through attention visualization [29]. These works collectively highlight the necessity of sophisticated spatiotemporal feature analysis in echocardiographic applications.

Despite their contributions to spatiotemporal feature processing, most existing approaches adopt a paradigm where spatial and temporal features are extracted sequentially. This cascaded approach may limit the potential synergy between spatial and temporal characteristics that are inherently intertwined in cardiac motion. Addressing this limitation requires innovative architectural designs that can effectively handle both spatial and temporal features simultaneously, particularly in the context of deep learning frameworks.

2.3. Modern deep learning approaches for cardiac video analysis

Early works in cardiac imaging analysis demonstrated promising results even with static approaches. Zhang et al. showed the feasibility of automated echocardiogram interpretation using conventional CNNs [4], though limited to frame-by-frame analysis. More recent studies have explicitly addressed the temporal aspects of cardiac imaging [30,31], indicating how architectural innovations can enhance dynamic quality assessment tasks.

Notably, the R(2 + 1)D architecture introduced an important concept by decomposing 3D convolutions into separate spatial and temporal components [32]. While this architectural design principle was originally developed for action recognition, it offers valuable insights for medical video processing, where precise temporal relationships are crucial for diagnosis. On the other hand, emerging developments in CNN architectures, specifically ConvNeXt [33], have introduced critical innovations in feature processing through a pure convolution-based approach. The incorporation of larger kernel sizes and layer normalization in ConvNeXt effectively addresses the limitations of traditional CNNs while maintaining computational efficiency. Moreover, through depthwise convolution and inverted bottleneck structures, ConvNeXt combines the advantages of inductive bias in CNNs with transformer-like global attention mechanisms, making such architectural designs markedly suitable for medical image analysis, where both local and global features are vital. Furthermore, the Global Response Normalization (GRN) mechanism introduced in ConvNeXt-V2 notably addresses the feature collapse problem by enhancing feature contrast and

maintaining diverse channel responses [34].

In summary, while significant progress has been made in automated echocardiographic analysis, several critical challenges remain unaddressed. First, existing approaches often treat spatial and temporal feature processing as separate steps, potentially missing crucial spatio-temporal interactions intertwined in cardiac motion patterns. Second, while recent architectural innovations like R(2 + 1)D and ConvNeXt have shown promise in their respective domains, their direct application to echocardiographic quality assessment is limited by the unique characteristics of cardiac imaging, such as the need to maintain temporal coherence while handling varying image quality. Third, current methods rarely consider the specific hallmarks of echocardiographic imaging in their architectural design, potentially limiting their effectiveness in clinical applications.

These observations collectively highlight the need for a specialized architecture that can: 1) simultaneously process spatial and temporal features while maintaining their interdependencies, 2) adapt and enhance existing architectures to effectively handle the dynamic nature of cardiac ultrasound, and 3) incorporate echocardiographic-specific hallmarks into the architectural design to enhance clinical effectiveness. These challenges motivate our development of the ST-R(2 + 1)D-ConvNeXt architecture, which specifically addresses these limitations through integrated spatiotemporal feature disentanglement and echocardiogram-specific design considerations.

3. Materials

The dataset comprises 514 labeled echocardiogram videos from 514 different patients collected at National Taiwan University Hospital (NTUH) between January 2019 and December 2022, acquired using Philips iE33 and EPIQ 7C imaging systems. Fig. 2 shows the temporal sequences of echocardiogram frames: the first two rows display high-quality images, and the last three rows demonstrate different quality problems such as a missing left ventricular wall (IQ0097), noise (IQ0009), and excessive degradation where LV contours are barely recognizable (IQ0209). For model development and evaluation, we randomly split the dataset into two parts: 438 videos for training and validation, and 76 videos for independent testing. The training-validation set (438 videos) was further divided using 5-fold cross-validation, resulting in 351 videos for training and 87 for validation in each fold. Finally, each frame was resized to 112×112 pixels after cropping and de-identification to protect patient privacy.

To the best of our knowledge, there is no universally accepted standard for assessing echocardiogram quality. Despite various proposed methods [35,36], their ability to reflect clinically vital features, which can be utilized by doctors, remains uncertain. To this end, we have collaborated with NTUH cardiologists to annotate data, ensuring that this study aligns with the actual needs for precise medical assessment.

Although GLS is crucial for diagnosis [27], a single GLS value alone may not particularly provide meaningful insights regarding quality assessment. Moreover, the image quality could potentially limit the reliability of GLS [6]. Instead of relying on direct annotations, our method evaluates the quality of echocardiograms by comparing GLS discrepancies between annotations from cardiologists and from TomTec, an automated software designed for analyzing medical images. We quantify the difference *dif* between two GLS values from manual cardiologist assessments y_a and the automated annotation y_m by the formula:

$$dif = \left| \frac{y_a - y_m}{y_m} \right| \times 100\% \quad (1)$$

We then apply a typical 10 % threshold, recommended by [37] for LVEF, to the calculated *dif* to define echocardiogram quality, re-formulating the task as a binary classification problem. An echocardiogram is

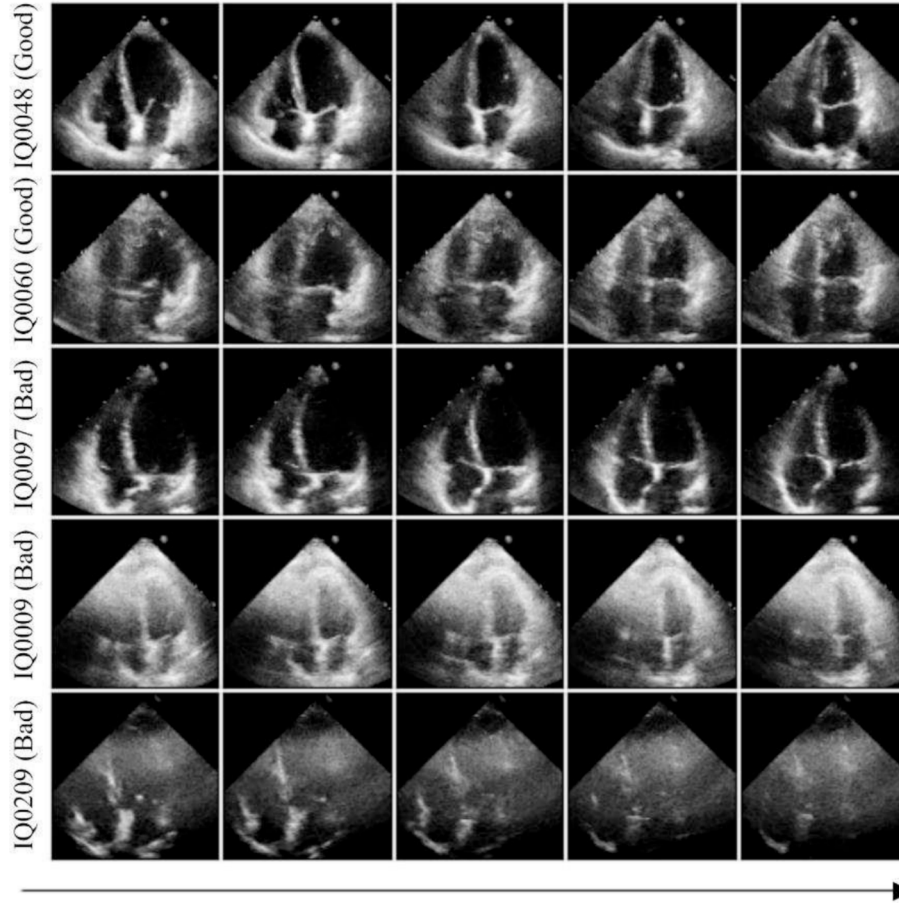


Fig. 2. Representative examples from our echocardiogram dataset. Each row shows a temporal sequence of continuous frames from an echocardiogram. The first two rows demonstrate high-quality images with clear cardiac structures and well-defined chamber walls. The last three rows display various quality issues: IQ0097 shows a missing left ventricular wall; IQ0009 exhibits significant noise artifacts; and IQ0209 displays excessive degradation where left ventricular (LV) contours are barely recognizable.

considered low-quality (positive) if the difference exceeds 10 %; and high-quality (negative) otherwise. This approach, relying on the variance between manual and automated GLS annotations, establishes a consensus-based evaluation.

4. Methods

To address echocardiographic quality surveillance, we propose a two-stage method consisting of a pre-processing stage for LV

segmentation and a quality prediction stage focused on the LV volume, as illustrated in Fig. 3. This approach emphasizes the importance of LV chambers in quality assessment. This systematic approach allows for focused analysis of the left ventricular region while reducing computational complexity. The following sections detail each stage of this framework.

4.1. Left ventricular VOI acquirement

In the pre-processing stage, we leverage the left ventricular segmentation model branch from EchoNet [38] to identify volumes of interest (VOIs) by training on the EchoNet-Dynamic dataset, which includes 10,036 cardiac echo-videos with labels like LV volumes at end-systole and end-diastole [39]. After training, we apply this model to segment our dataset and obtain the VOIs, focusing on the LVs. We then resize these frames to 64×64 pixels and prepare them for the quality prediction stage. Note that we conservatively enlarge measured VOIs to preserve LV walls, which provide critical information for analysis [40]. Fig. 4 shows the pre-processing flow. For simplicity, the figure shows only one frame as an example. Note that the blue region in the figure is merely for demonstrating the segmentation result and is not exploited in our model.

This strategy of isolating LV volumes enhances our approach, yielding several benefits: 1) The quality prediction model can exclusively analyze LV regions, aligning with cardiological expertise and streamlining the process. 2) The model performance is improved. 3) The computational costs in both time and memory are greatly reduced, as

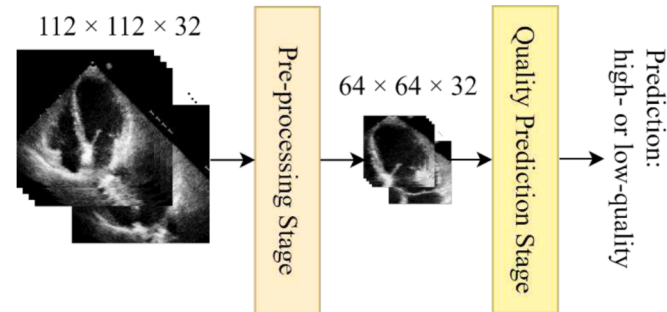


Fig. 3. Overview of the proposed two-stage framework for echocardiogram quality surveillance. The input video consists of 32 frames, each with a spatial resolution of 112×112 pixels. The pre-processing stage extracts the left ventricular volume, which is cropped and resized to 64×64 pixels per frame for the subsequent quality prediction stage.

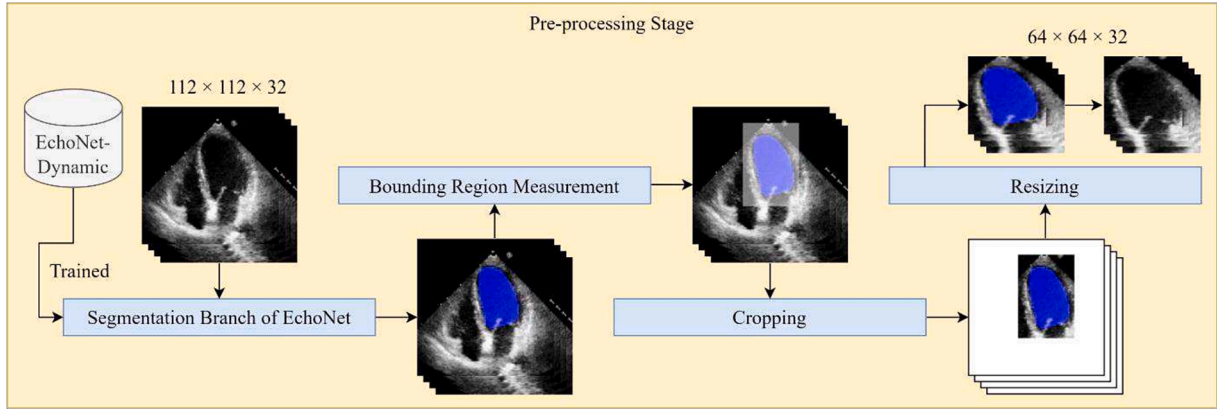


Fig. 4. The flow of the pre-processing stage for left ventricular (LV) volume of interest (VOI) acquisition. The process includes: 1) original input video sequence ($112 \times 112 \times 32$ pixels); 2) LV segmentation using EchoNet model (blue region indicates segmented area); and 3) final cropped VOI ($64 \times 64 \times 32$ pixels) used for quality assessment. Note that while the figure demonstrates a few frames for visualization purposes, the actual processing pipeline operates on 32 consecutive frames.

frames are downsized from 112×112 pixels to 64×64 pixels.

4.2. Quality prediction

In the quality prediction stage, we propose *ST-R(2 + 1)D-ConvNeXt*, a novel ConvNeXt-based [33,34] model, designed to analyze echocardiographic hallmarks and disentangle spatiotemporal features. This model takes well-prepared videos in the size of $64 \times 64 \times 32$ frames and predicts a binary output (high- or low-quality).

4.2.1. $R(2 + 1)D$ architecture

Given that ConvNeXt (V1 and V2) [33,34], which is optimized for image data and not videos, we integrate the renowned $R(2 + 1)D$ architecture [32]. The $R(2 + 1)D$ layer consists of a 2D convolution to process spatial dimensions, followed by a 1D convolution to handle the temporal dimension. This factorization allows for a more precise exploration of video characteristics, overcoming the problem that 3D convolution lacks the capacity to properly address different feature

types in videos (spatial and temporal).

In our design, each ConvNeXt block is tailored by replacing the original 2D convolution layer (as the depth-wise convolution layer) with an $R(2 + 1)D$ layer. We refer to this modification as *R(2 + 1)D-ConvNeXt*. This block can profit from both the excellent potency of ConvNeXt and the space-time disentanglement ability embedded in $R(2 + 1)D$.

4.2.2. Squeeze-and-excitation block

The Squeeze-and-Excitation (SE) block, initially designed to improve channel-wise interactions for 2D images [41], as depicted in Fig. 5 (a), faces challenges when adapted directly to video data due to spatio-temporal conflicts [32]. A slightly more complex approach is firmly desired for video data.

Moreover, in echocardiography, the position of the heart remains fixed during imaging as probes are typically stationary, leading to minimal spatial changes across frames [25,42]. Consequently, the effectuality in fusing echocardiographic features relies more on long-

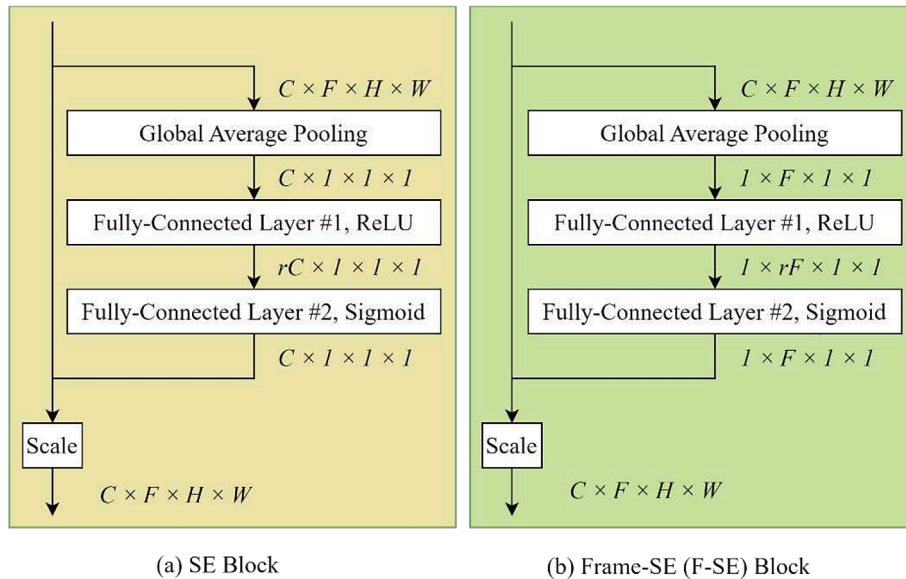


Fig. 5. Architecture comparison of attention blocks: (a) Squeeze-and-Excitation (SE) block for channel-wise feature recalibration, and (b) Frame-wise Squeeze-and-Excitation (F-SE) block for temporal feature competition. C , F , H , and W denote channel, temporal (frame), height, and width dimensions respectively; r represents the reduction ratio that determines the bottleneck size in attention modules. The F-SE block is specifically designed to enhance model capability in capturing temporal dynamics in videos through direct competition among features from different frames. This mechanism is crucial for assessing motion patterns and temporal consistency.

distance temporal information than on spatial relationships alone. This negates the stringent need for spatial-wise attention. In addition, we experimentally show that dissociating space and time is more crucial than focusing on spatial-wise feature interactions.

To this end, as shown in Fig. 5 (b), we introduce the F-SE block to intensify feature competition and contrast across frames, promoting well-distributed attention within the model. Each F-SE block is paired with an SE block to form an FC-SE block, where 'FC' stands for frame-and-channel. This combination harnesses channel-wise attention to better utilize features. The FC-SE block is specifically designed to focus on the temporal axis by distributing attention across both channels and time, aligning with echocardiographic hallmarks.

4.2.3. Global response normalization block

In ConvNeXt-V2, the Global Response Normalization (GRN) block addresses variations in channel responsiveness and prevents the feature collapse problem observed in ConvNeXt-V1 [33,34]. GRN for 2D data has three steps: spatial feature aggregation via L2-norm pooling, channel-wise normalization to enrich contrast across channels, and output calibration through element-wise multiplication with the input. A residual path is applied to preserve information from the origin.

One simple way to apply GRN to video data is to simply add an additional dimension to the GRN block and the rest remains unchanged. This effortless approach may improperly handle representations in videos when applied in such 3D manners. To this end, we introduce Spatiotemporal GRN (ST-GRN) for echocardiogram analysis. This refinement, based on GRN, additionally emphasizes the intrinsic dissimilarities between space and time within video data.

ST-GRN retains the core steps of the original GRN, namely, feature aggregation, feature normalization, and feature calibration. However, ST-GRN presents two symmetric branches that optimize feature responses separately for spatial and temporal axes. In overview, ST-GRN has four main steps: feature aggregation, feature normalization, cross-dimension re-normalization, and feature calibration.

Given an input $X \in \mathcal{R}^{F \times H \times W \times C}$, where F , H , W , and C respectively represent the temporal dimension, height, width, and channels, X is processed separately through two L2-norm-based aggregation functions: a spatial function $\mathcal{S}$ and a temporal function $\mathcal{T}$. The spatial function, $\mathcal{S}$, can merge spatial information, and the temporal function, $\mathcal{T}$, fuses features along the temporal axis. These operations produce two elements for each channel i , $X_i^S \in \mathcal{R}^F$ and $X_i^T \in \mathcal{R}^{H \times W}$, where $1 \leq i \leq C$.

$$\mathcal{S}(X) = X^S := \mathcal{R}^{F \times H \times W \times C} \rightarrow \mathcal{R}^{F \times C} \quad (2)$$

$$\mathcal{T}(X) = X^T := \mathcal{R}^{F \times H \times W \times C} \rightarrow \mathcal{R}^{H \times W \times C} \quad (3)$$

Building on the original GRN, the feature normalization function $\mathcal{N}$ in ST-GRN divides each channel element by the channel mean. This operation is, again, applied symmetrically to spatial $X^S = \{X_1^S, X_2^S, \dots, X_C^S\}$ and temporal $X^T = \{X_1^T, X_2^T, \dots, X_C^T\}$ components:

$$M^S = \frac{\sum_{i=1}^C X_i^S}{C}, M^T = \frac{\sum_{i=1}^C X_i^T}{C} \quad (4)$$

$$\mathcal{N}_i(X^S) = \frac{X_i^S}{M^S}, \mathcal{N}_i(X^T) = \frac{X_i^T}{M^T} \quad (5)$$

where $M^S \in \mathcal{R}^F$ and $M^T \in \mathcal{R}^{H \times W}$ are the means of X^S and X^T over channels, respectively.

We further denote $X^{Ns} \in \mathcal{R}^{F \times C}$ as the output of $\mathcal{N}(X^S) = \{\mathcal{N}_1(X^S), \mathcal{N}_2(X^S), \dots, \mathcal{N}_C(X^S)\}$ and $X^{Nr} \in \mathcal{R}^{H \times W \times C}$ as the output of $\mathcal{N}(X^T)$. Since the shapes of X_i^S and M^S are the same (same with X_i^T and M^T), the element-wise division in $\mathcal{N}_i$ produces the same-shape output as well. These operations are coupled to enrich channel-wise contrast along the spatial and temporal dimensions, respectively.

Atop that, different from the original GRN, we apply the cross-

dimension re-normalization function $\mathcal{D}$. Similar to the previous feature normalization step, we calculate the means but, at this time, across space and time:

$$M^{Ns} = \frac{\sum_{j=1}^F X_j^{Ns}}{F}, M^{Nr} = \frac{\sum_{k=1}^{H \cdot W} X_k^{Nr}}{H \cdot W} \quad (6)$$

where $M^{Ns} \in \mathcal{R}^C$ and $M^{Nr} \in \mathcal{R}^C$ indicate the mean values of X^{Ns} and X^{Nr} at spatial and temporal dimensions.

Following that, a pair of cross-dimension re-normalization operations is performed:

$$\mathcal{D}_j(X^{Ns}) = \frac{X_j^{Ns}}{M^{Nr}}, \mathcal{D}_k(X^{Nr}) = \frac{X_k^{Nr}}{M^{Ns}} \quad (7)$$

where $1 \leq j \leq F$ and $1 \leq k \leq (H \cdot W)$.

Due to the previously unaddressed interaction between spatial and temporal dimensions in the ST-GRN block, we develop this re-normalization step that compensates by adjusting features across both dimensions simultaneously, enhancing feature co-inhibition. This secondary normalization is akin to the earlier channel-wise normalization, but spatially and temporally. The paired results are denoted as $X^{Ds} = \mathcal{D}(X^{Ns}) \in \mathcal{R}^{F \times C}$ and $X^{Dr} = \mathcal{D}(X^{Nr}) \in \mathcal{R}^{H \times W \times C}$.

Lastly, in ST-GRN, feature calibration incorporates two learnable parameters, γ and β , and a residual path to generate the final result of the ST-GRN block Y :

$$Y = X + (\gamma * (X * X^{Ds} * X^{Dr}) + \beta) \in \mathcal{R}^{F \times H \times W \times C} \quad (8)$$

4.2.4. ST-R(2 + 1)D-ConvNeXt

The final ST-R(2 + 1)D-ConvNeXt module is illustrated in Fig. 6. The figure summarizes the integration of modified components from R(2 + 1)D [32], SE block [41], ConvNeXt-V1 [33], and GRN from ConvNeXt-V2 [34]. By combining these modified architectures and blocks, we finalize ST-R(2 + 1)D-ConvNeXt, an effective deep learning model for echocardiogram quality surveillance, specifically designed considering echocardiographic hallmarks and spatiotemporal feature disentanglement.

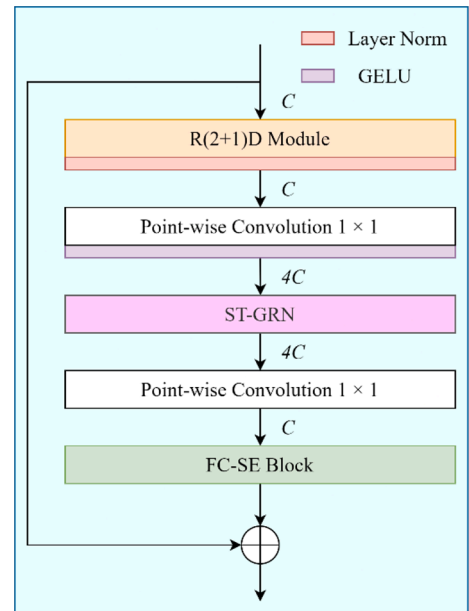


Fig. 6. The ST-R(2 + 1)D-ConvNeXt module architecture, where ST stands for spatiotemporal. The module integrates R(2 + 1)D convolution for spatiotemporal feature disentanglement, Spatiotemporal Global Response Normalization (ST-GRN) for feature calibration, and Frame-Channel Squeeze-and-Excitation (FC-SE) block for attention mechanism.

Moreover, Fig. 7 reveals the structure of ST-R(2 + 1)D-ConvNeXt, which has a stem and four main stages, interlaced by down-sampling layers. Each stage repeats the ST-R(2 + 1)D-ConvNeXt module a certain number of times. Following the fourth stage, an average pooling layer with a set of fully-connected layers is employed to generate the final prediction of echocardiogram quality.

5. Experiment

In this section, we conduct comprehensive experiments to validate the effectiveness of the proposed ST-R(2 + 1)D-ConvNeXt model through ablation studies and comparisons with state-of-the-art architectures. Moreover, we also examine key model components, including attention mechanisms and GRN blocks, with detailed analysis of their individual and combined contributions to echocardiographic quality assessment.

5.1. Experimental setup and training details

In experiments, we validate model capabilities to accurately predict the quality of echocardiograms by comparing different architectural components, including types of convolution layers, attention mechanisms, and GRN blocks. The primary training hyperparameters are shown in Table 1.

All experiments are conducted in the same 5-fold cross-validation workflow. Key model specifications include the number of repetitions in each stage $\langle N_1, N_2, N_3, N_4 \rangle$ is set to $\langle 2, 2, 6, 2 \rangle$ and the number of channels in each stage is respectively 16, 32, 64, and 128. Besides, we use a Conv3D with a kernel size of 2 and a stride of 2 in all three dimensions as the down-sampling layer. Frame-Channel Squeeze-and-Excitation (FC-SE) blocks were incorporated with a reduction ratio of 2, applied across all four stages. For the video pre-processing, each video clip contained 32 consecutive frames. During training, we applied data augmentation, while for validation and testing, we employed test-time augmentation, sampling 32 continuous frames with a 16-frame step between samples and feeding each 32-frame segment into the model. We then use a logit-ensemble method, averaging all logits to produce the final output. All implementations were carried out using PyTorch on an NVIDIA GeForce RTX 3090 GPU. The probability threshold for binary classification was set at 0.5.

As for performance metrics, given the critical nature of echocardiographic quality assessment in clinical settings, we employ comprehensive evaluation metrics to assess the model performance. Sensitivity and specificity are particularly crucial as they respectively measure model

Table 1

Primary training hyperparameters for the proposed ST-R(2 + 1)D-ConvNeXt model.

Argument	Value	Note
# Epochs	300	Early-stopping applied (patience = 50)
Batch size	100	
Loss function	Binary cross-entropy loss	R^2 weighting applied
Optimizer	Adam [43]	No weight-decaying
Learning rate (LR)	10^{-4}	No LR scheduling
Data augmentation	Random elastic deformation + random rotation from -60° to 60°	

capabilities to identify low-quality videos that might lead to misdiagnosis (sensitivity) and to confirm high-quality ones that ensure reliable clinical assessments (specificity). Positive predictive value (PPV) and negative predictive value (NPV) are included to evaluate the trustworthiness of model predictions in real-world scenarios, where the cost of misclassifying a low-quality video as high-quality could impact clinical decisions. The F1-score provides a balanced performance measure, particularly important given the potential imbalance between high- and low-quality cases in clinical practice. For statistical analysis, McNemar's test was employed to calculate p-values to assess the statistical significance of differences between model performances. The test was implemented using the mlxtend library (version 0.22.0) in Python. In our terminology, 'positive' indicates low quality, and 'negative' refers to high quality. For example, true positive (TP) means the ground truth is low-quality and the model predicts correctly.

To validate the training stability and potential overfitting of the proposed ST-R(2 + 1)D-ConvNeXt model, we examine the training and validation curves across the 5-fold cross-validation experiments, as shown in Fig. 8. The loss curves demonstrate consistent convergence behavior, with the training loss stabilizing between 0.5 and 0.6 and the validation loss around 0.6 to 0.7. The gap between training and validation metrics (approximately 15 to 20 % for loss and 5 to 10 % for accuracy) remains within acceptable ranges throughout the training process, indicating no significant overfitting.

5.2. Ablation study

We begin the experiment section with an ablation study to assess the

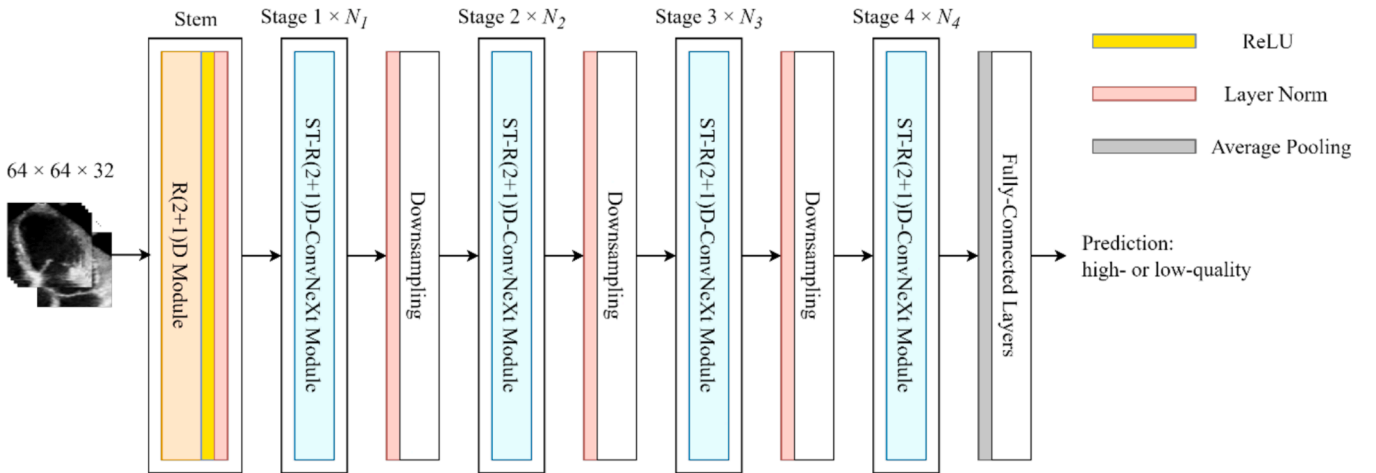


Fig. 7. The overall architecture of the proposed ST-R(2 + 1)D-ConvNeXt model for left ventricular echocardiogram quality surveillance. The model consists of a stem layer, four main stages with ST-R(2 + 1)D-ConvNeXt modules (number of modules per stage: 2, 2, 6, 2, respectively), down-sampling layers, and final classification layers. Input size: $64 \times 64 \times 32$ (height $\times$ width $\times$ frames).

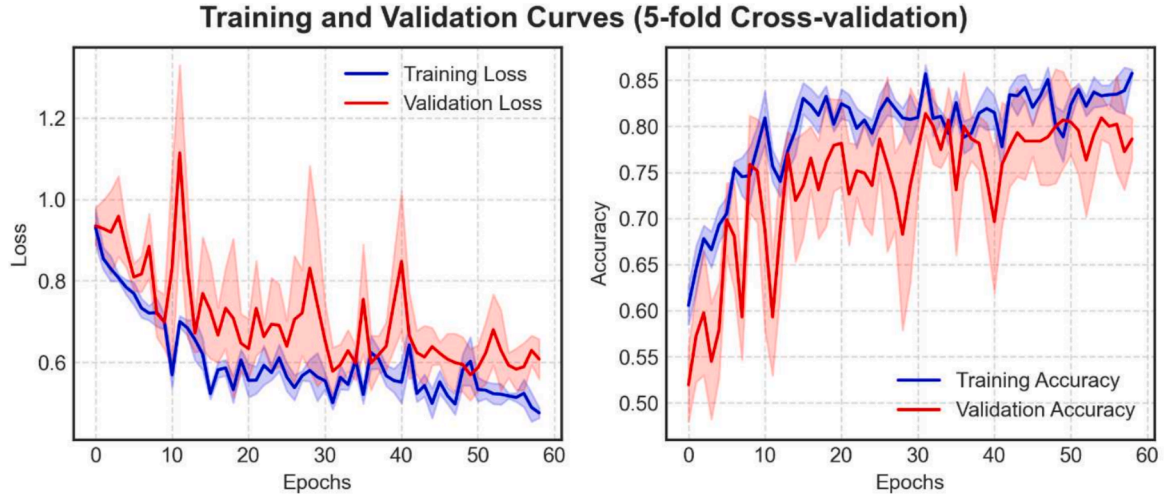


Fig. 8. Training and validation curves of the proposed ST-R(2 + 1)D-ConvNeXt model with 5-fold cross-validation. Left: loss curves showing the binary cross-entropy loss of training and validation sets. Right: accuracy curves showing the classification performance on training and validation sets. The solid lines represent mean values while the shaded areas indicate standard deviation across folds. All curves are truncated at epoch 57, which represents the earliest stopping point among all folds, to ensure fair visualization of the cross-validation results.

efficacy of the proposed components in a hierarchical manner. Our study starts with a 3D adaptation of ConvNeXt-V2 [34], where all 2D convolution layers are replaced with 3D layers to handle non-2D data. We then integrate the R(2 + 1)D architecture, which serves as the baseline for subsequent experiments due to its fundamental role in spatiotemporal feature disentanglement. Building upon this R(2 + 1)D-ConvNeXt baseline, we examine the individual contributions of FC-SE and ST-GRN modules, which are specifically designed to leverage these disentangled features, making their evaluation in isolation methodologically inappropriate. This architectural interdependence reflects our fundamental design philosophy of leveraging the unique temporal characteristics of echocardiographic sequences, rather than treating them as generic 3D volumes. The complete experimental configurations and results are detailed in Table 2.

The results by and large exhibit an upward trend in accuracy and AUC, indicating the positive impact of each component. Specifically, incorporating R(2 + 1)D improves accuracy by a substantial gap of 5.00 %, underscoring its effectiveness in spatiotemporal feature disentanglement. Remarkably, while the 3D ConvNeXt-V2 baseline shows a significant imbalance in predictions, with high specificity (87.03 %) but poor sensitivity (63.08 %), the incorporation of R(2 + 1)D helps achieve a more balanced performance. Although this initially results in a lower specificity (75.14 %) and PPV (78.64 %), the R(2 + 1)D architecture indeed addresses the severe prediction bias present in 3D ConvNeXt, leading to more balanced classification capabilities.

Building upon this R(2 + 1)D-based framework, we examined the

individual contributions of FC-SE and ST-GRN modules. The addition of FC-SE to the baseline notably improves specificity from 75.14 % to 81.08 % and PPV from 78.64 % to 82.18 %, though with a slight decrease in sensitivity (79.49 %). Meanwhile, incorporating ST-GRN alone achieves the highest AUC (0.89) while maintaining high sensitivity (80.51 %), despite a marginal decrease in overall accuracy. These complementary characteristics of FC-SE (specificity enhancement) and ST-GRN (boundary optimization) motivated their integration in the final architecture.

With the integration of both FC-SE and ST-GRN, the proposed ST-R(2 + 1)D-ConvNeXt is fully realized and demonstrates significant performance improvements. Notably, compared with the baseline with FC-SE, accuracy increases to 82.63 % (+2.37 %), PPV to 82.44 % (+0.26 %), NPV to 82.92 % (+3.86 %), F1-score to 83.24 % (+2.66 %), and AUC to 0.89 (+0.01). Key enhancements are observed in accuracy, NPV, and F1-score. Worthy of discussion, the reduction of false positives and false negatives in an equilibrating manner re-achieves a satisfactory sensitivity of 84.10 %. The ST-GRN block effectively counteracts the potential decreases in sensitivity introduced by FC-SE, demonstrating how the complementary nature of these components yields optimal performance across all metrics.

5.3. Comparison of model architectures

We opine that using only 3D convolution layers is not sufficient to fully decipher the spatiotemporal features of echocardiographic data. To

Table 2

Ablation study of the proposed ST-R(2 + 1)D-ConvNeXt architecture. The study begins with a 3D adaptation of ConvNeXt-V2, followed by the integration of R(2 + 1)D design as our architectural foundation. This R(2 + 1)D-integrated architecture serves as our baseline for further experiments. Building upon this R(2 + 1)D-based framework, we examine the contributions of FC-SE and ST-GRN components, which are specifically designed to leverage spatiotemporal features and echocardiographic hallmarks.

	Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	F1-score (%)	AUC	p-value
ConvNeXt-V2 (3D model)	74.74 ± 3.67	63.08 ± 8.36	87.03 ± 7.91	84.75 ± 7.44	69.41 ± 3.81	71.69 ± 5.25	0.87	0.016*
+ R(2+1)D (baseline)	79.74 ± 2.44	84.10 ± 4.97	75.14 ± 8.61	78.64 ± 4.92	82.20 ± 3.64	81.01 ± 1.70	0.87	0.046*
Baseline with FC-SE	80.26 ± 2.20	79.49 ± 3.24	81.08 ± 7.83	82.18 ± 5.53	79.06 ± 1.09	80.58 ± 1.12	0.88	0.023*
Baseline with ST-GRN	78.16 ± 2.95	80.51 ± 3.48	75.68 ± 7.25	78.11 ± 4.89	78.74 ± 2.37	79.13 ± 2.38	0.89	0.027*
Proposed	82.63 ± 1.75	84.10 ± 2.51	81.08 ± 2.42	82.44 ± 1.90	82.92 ± 2.31	83.24 ± 1.73	0.89	—

* denotes p-value < 0.05.

validate this, we demonstrate the difference between model architectures and the use of $R(2 + 1)D$ architecture [32] and 3D convolution layers. We adopt ResNet [44], ConvNeXt-V2 [33,34], and InceptionNeXt [45] as our baseline models, representing the evolution of pure CNN-based architectures from pre-transformer era to contemporary designs.

Table 3 reports the performance of different architectures. The superiority of $R(2 + 1)D$ over 3D convolution is consistently validated across all models, where $R(2 + 1)D$ ResNet-38 outperforms its 3D counterpart by 3.16 %, $R(2 + 1)D$ ConvNeXt-V2 shows an even larger improvement of 5.00 %, and $R(2 + 1)D$ InceptionNeXt surpasses 3D InceptionNeXt by 3.95 %. This consistent pattern strongly suggests that standard 3D convolution fails to effectively extract informative features from echocardiographic videos.

We further incorporate TAdaConvNeXt, which is a ConvNeXt-based model specifically tailored for video understanding with temporally adaptive convolutions [46]. The outstanding performance of TAdaConvNeXt with an accuracy of 79.21 % validates our hypothesis about the importance of proper temporal feature handling in echocardiographic analysis. TAdaConvNeXt achieves relatively balanced performance across all metrics, for instance, a sensitivity of 80.51 % and a specificity of 77.84 %, demonstrating its robust feature extraction capability.

Notably, both 3D InceptionNeXt (71.84 %, in terms of accuracy) and $R(2 + 1)D$ InceptionNeXt (75.79 %) present poor performance compared to the configurations using ConvNeXt. In our opinion, this is likely because InceptionNeXt was originally designed to decompose large kernels for efficiency in 2D image processing tasks, where its parallel branch structure effectively reduces computation costs while maintaining spatial feature extraction capability. However, in the context of echocardiographic video analysis where temporal dynamics play a crucial role, this decomposition strategy may not be optimal for capturing the complex spatiotemporal relationships inherent in cardiac motion.

Shifting to the comparison between our proposed ST- $R(2 + 1)D$ -ConvNeXt and these state-of-the-art architectures, we observe superior performance in most metrics, achieving the highest accuracy (82.63 %), sensitivity (84.10 %, tied with $R(2 + 1)D$ ConvNeXt-V2), NPV (82.92 %), F1-score (83.24 %), and AUC (0.89). While 3D ConvNeXt-V2 shows radically high specificity (87.03 %) and PPV (84.75 %), this comes at the cost of significantly lower sensitivity (63.08 %) and NPV (69.41 %), indicating a severely biased model behavior and an unwanted imbalance. In contrast, our model maintains balanced performance across thorough metrics, with sensitivity, specificity, PPV, and NPV all exceeding 80 %, displaying robustness and reliability. The fact that $R(2 + 1)D$ ConvNeXt-V2 falls short in specificity further highlights the overall superiority of the proposed model.

5.4. Differences in attention mechanisms

To assess the effectiveness of FC-SE, we compare various attention

mechanisms, including SE [41], F-SE, CF-SE, and FC-SE, where 'CF' denotes SE followed by an F-SE, and 'FC' refers to F-SE followed by an SE. All of them focus on channel- and/or frame-wise feature interactions.

Results from Table 4 indicate that using either SE or F-SE alone leads to a noticeable performance deficiency, stressing the necessity for both channel-wise and frame-wise attention in echocardiographic analysis. Nonetheless, F-SE outshines SE by a pronounced margin, increasing accuracy by 2.11 %. These observations experimentally support the importance of frame-wise attention as discussed earlier.

5.5. Effect on ST-GRN

To advocate ST-GRN, we experiment with various GRN blocks: No GRN, 3D-GRN, and ST-GRN, where, for clarification, 3D-GRN refers to the original GRN but all operations are straightforwardly replaced by their 3D counterparts. The results are reported in Table 5.

The results show that 3D-GRN has significantly bad performance scores, including the lowest accuracy (77.89 %), specificity (73.51 %), PPV (77.47 %), and F1-score (79.24 %). The difference in specificity between 3D-GRN and No GRN is 7.57 %, suggesting that 3D-GRN is an over-facile design for handling videos due to the absence of spatiotemporal feature disentanglement. Compared to No GRN, 3D-GRN decreases accuracy by 2.37 %, confirming the detrimental impact of 3D-GRN. In our opinion, this is because 3D-GRN improperly treats all axes as spatial ones and hence the temporal features cannot be beneficially utilized. This counterfactual treatment on the temporal axis provokes a high False Positive (FP) rate and, consequently, an unacceptably low specificity. Last but not least, this experiment proves that the presented ST-GRN design conforms to the "rules" of handling space and time axes, demonstrating an exceptional capability for video processing. This is recognizable in accuracy by an improving gap of 4.74 % compared to 3D-GRN.

6. Discussion

We discuss three critical aspects of our work: 1) the technical advantages of left ventricular VOI acquirement, 2) the fundamental characteristics that guide our model design, and 3) the practical implications in clinical settings. These discussions not only validate our methodological choices but also outline the clinical viability of our approach in real-world applications.

6.1. Left ventricular VOI acquirement

In this paper, we introduce a two-stage method and a novel deep learning model tailored for echocardiogram quality surveillance. The first stage focuses on acquiring left ventricular VOIs. Removing non-LV parts ensures that the quality prediction aligns with cardiologists and effectuates an elegant algorithm. Additionally, the computational cost

Table 3

Performance comparison of CNN-based architectures with 3D, $R(2 + 1)D$, and temporally adaptive convolutions for echocardiogram quality assessment. 3D and $R(2 + 1)D$ variants denote direct modifications from original 2D models, where ResNet-38, ConvNeXt-V2, and InceptionNeXt represent architectures from pre-transformer era to contemporary designs. TAdaConvNeXt is a dedicated video-oriented ConvNeXt-based model with temporally adaptive convolutions.

	Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	F1-score (%)	AUC	p-value
3D ResNet-38	70.79 ± 9.79	83.08 ± 8.36	57.84 ± 28.33	70.89 ± 11.99	75.73 ± 2.91	75.12 ± 4.43	0.81	< 0.001*
$R(2 + 1)D$ ResNet-38	73.95 ± 0.53	68.72 ± 4.10	79.46 ± 4.39	78.15 ± 2.63	70.79 ± 1.64	72.97 ± 1.28	0.83	< 0.001*
3D ConvNeXt-V2	74.74 ± 3.67	63.08 ± 8.36	87.03 ± 7.91	84.75 ± 7.44	69.41 ± 3.81	71.69 ± 5.25	0.87	0.016*
$R(2 + 1)D$ ConvNeXt-V2	79.74 ± 2.44	84.10 ± 4.97	75.14 ± 8.61	78.64 ± 4.92	82.20 ± 3.64	81.01 ± 1.70	0.87	0.046*
3D Inception-NeXt	71.84 ± 11.58	75.90 ± 6.61	67.57 ± 17.93	72.34 ± 11.21	71.19 ± 12.44	73.89 ± 8.95	0.80	< 0.001*
$R(2 + 1)D$ Inception-NeXt	75.79 ± 2.95	74.87 ± 6.15	76.76 ± 6.53	77.56 ± 4.05	74.65 ± 3.75	75.97 ± 3.36	0.86	0.010*
TAdaConvNeXt	79.21 ± 2.41	80.51 ± 2.61	77.84 ± 6.71	79.63 ± 4.18	79.17 ± 1.42	79.94 ± 1.65	0.88	0.027*
Proposed	82.63 ± 1.75	84.10 ± 2.51	81.08 ± 2.42	82.44 ± 1.90	82.92 ± 2.31	83.24 ± 1.73	0.89	–

* denotes p-value < 0.05.

^ same as the first two cases in the ablation study.

Table 4

Comparison of attention mechanisms in the ST-R(2 + 1)D-ConvNeXt model. SE: Squeeze-and-Excitation; F-SE: Frame-wise SE; CF-SE: Channel-followed-by-Frame SE; FC-SE: Frame-followed-by-Channel SE.

Attention type	Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	F1-score (%)	AUC	p-value
SE	77.63 ± 1.44	80.51 ± 1.26	74.59 ± 3.24	77.03 ± 2.18	78.41 ± 1.07	78.71 ± 1.10	0.88	0.041*
F-SE	79.74 ± 3.07	78.46 ± 6.80	81.08 ± 8.37	82.22 ± 6.61	78.52 ± 4.30	79.84 ± 3.02	0.88	0.041*
CF-SE	80.26 ± 1.18	81.54 ± 5.48	78.92 ± 6.92	80.81 ± 4.42	80.61 ± 3.16	80.88 ± 1.21	0.88	0.074
FC-SE	82.63 ± 1.75	84.10 ± 2.51	81.08 ± 2.42	82.44 ± 1.90	82.92 ± 2.31	83.24 ± 1.73	0.89	—

* denotes p-value < 0.05.

Table 5

Comparison of different Global Response Normalization (GRN) blocks in the ST-R(2 + 1)D-ConvNeXt model. No GRN: model without GRN block; 3D-GRN: conventional GRN extended to 3D; ST-GRN: proposed Spatiotemporal GRN.

GRN type	Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	F1-score (%)	AUC	p-value
No GRN	80.26 ± 2.20	79.49 ± 3.24	81.08 ± 7.83	82.18 ± 5.53	79.06 ± 1.09	80.58 ± 1.12	0.88	0.023*
3D-GRN	77.89 ± 2.68	82.05 ± 6.07	73.51 ± 10.86	77.47 ± 6.17	80.07 ± 3.20	79.24 ± 1.53	0.87	0.003*
ST-GRN	82.63 ± 1.75	84.10 ± 2.51	81.08 ± 2.42	82.44 ± 1.90	82.92 ± 2.31	83.24 ± 1.73	0.89	—

* denotes p-value < 0.05.

^ same as the third case in the ablation study.

can be greatly saved, in terms of both time and memory, as each 112×112 pixels frame is downsized to 64×64 pixels.

6.2. Spatiotemporal attributes and echocardiographic hallmarks

To develop an effective and efficient echocardiogram quality prediction model, we focus on two key data characteristics: spatiotemporal attributes and echocardiographic hallmarks. These foundations guide our research as we differentiate between images, videos, and echocardiograms. Videos differ from images by possessing spatiotemporality, which we address in the integration of R(2 + 1)D and ST-GRN to disentangle these representations. On the other hand, echocardiograms display unique traits like position consistency, topological homogeneity, and monochromaticity. These characteristics are crucial and explored in the FC-SE block, which focuses on enhancing feature competition across frames and channels, specifically excluding spatial interactions.

6.3. Medical applications

This work involves close collaboration with cardiologists, including annotations of GLS, concise establishments on LV VOI acquirement respecting the imaging protocol [47], and defect identification via explainable AI techniques. Specifically, we adopt Gradient-weighted Class Activation Mapping (Grad-CAM) [48], a prominent tool that visualizes regions influencing CNN decisions. Grad-CAM heatmaps offer valuable insights into multifaceted interpretations, directly supporting clinical decision-making by providing actionable feedback for image acquisition optimization. Based on this clinical integration, we adapt the idea of Grad-CAM to serve as a defect detector, enabling immediate quality assessment during image acquisition. Fig. 9 shows Grad-CAM visualizations across five models: ST-R(2 + 1)D-ConvNeXt, 3D ResNet, R(2 + 1)D ResNet, 3D ConvNeXt-V2, and R(2 + 1)D ConvNeXt-V2. We make some key observations:

- 1) Our model, ST-R(2 + 1)D-ConvNeXt, effectively identifies defects, particularly in low-quality cases, by detecting poorly positioned chambers (E), disappearing cardiac walls (G), and distorted contours (F).
- 2) The attention distribution of ST-R(2 + 1)D-ConvNeXt offers noticeably more relevant feature interpretation compared to other models. For instance, in (G), only ST-R(2 + 1)D-ConvNeXt precisely identifies the defect at the bottom-left region and correctly performs the prediction. This unique discriminative capability becomes even more pronounced in (C), where only our model remains unaffected by the irrelevant triangular void area at the upper-right corner.

- 3) Only ST-R(2 + 1)D-ConvNeXt properly maps precise LV contours in both high- and low-quality images, albeit in different ways. A good model may “pour” attention into the chamber interior when addressing high quality. In contrast, if the model seeks to identify something not interior, the irregular contours or inconsistent wall motions may be potential factors.
- 4) Despite similar rationales, ST-R(2 + 1)D-ConvNeXt outperforms 3D ConvNeXt-V2, possibly due to the inadequate handling of video data using 3D convolutions in latter one.

7. Limitations and future work

This study has several limitations that warrant discussion. First, despite our careful data collection process, the modest dataset size (514 videos) may limit the generalizability of the model across different clinical settings, particularly in hospitals with different imaging protocols. Second, the focus on A4C views, while justified for LV analysis, may exclude other important cardiac views that could provide complementary information for comprehensive cardiac assessment. Third, our current approach does not explicitly consider the periodic patterns in cardiac rhythm, which could provide valuable temporal information for quality assessment. Finally, although our model demonstrates promising results, the lack of external validation on independent datasets remains a significant limitation for clinical deployment.

To address these limitations, in the future, we aim to expand our dataset through multi-center collaboration with cardiologists, incorporating diverse patient populations and imaging protocols. We plan to implement a semi-automated annotation workflow, where our model generates initial annotations for expert refinement, thereby accelerating the data collection process while maintaining annotation quality. Additionally, we intend to further explore periodic patterns in cardiac rhythm within echocardiographic videos, another echocardiographic characteristic that inherently impacts cardiac function and assessment but has not been considered in this study. This exploration of cardiac periodicity could enhance our understanding of cardiac morphology and potentially advance echocardiographic analysis techniques by incorporating these intrinsic temporal patterns.

8. Conclusion

This paper presents a novel two-stage deep learning approach for dynamic LV echocardiography quality surveillance, re-formulating the task as a binary classification problem with expert cardiologist engagement. The first stage acquires left ventricular VOIs from dynamic echocardiograms and the second stage predicts quality by utilizing the

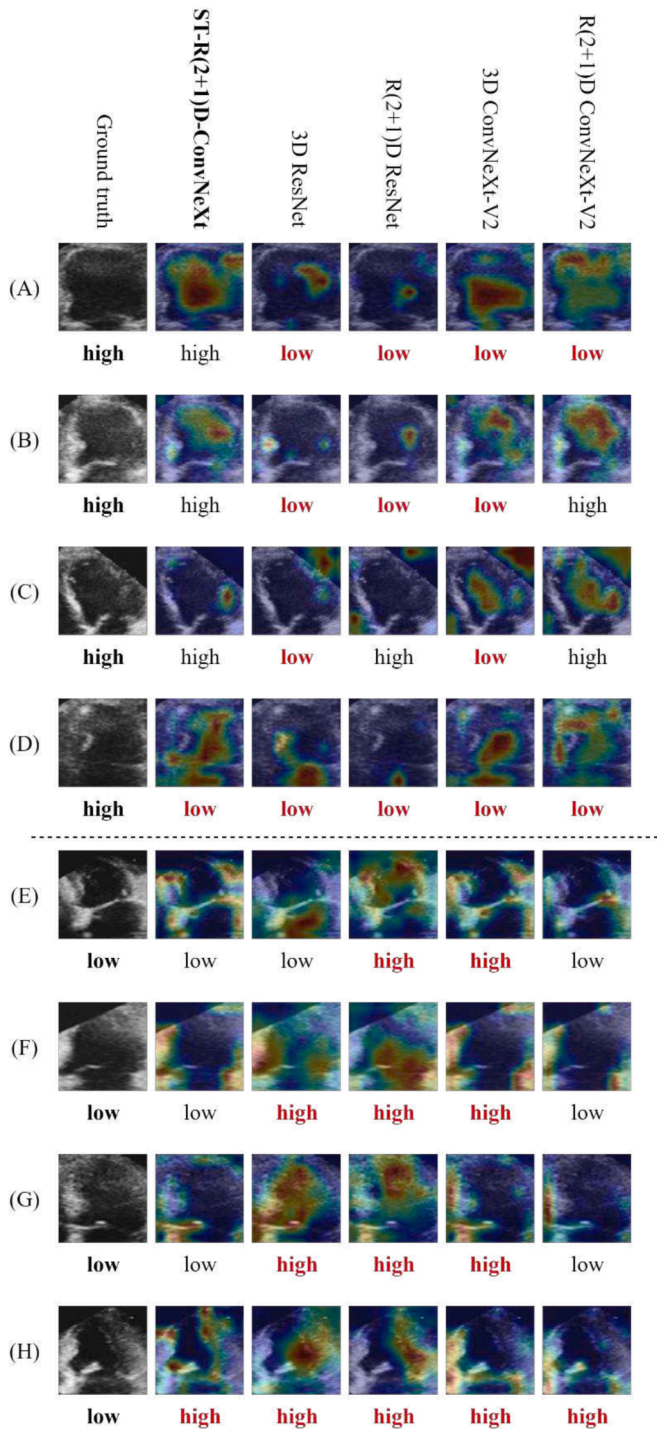


Fig. 9. Gradient-weighted Class Activation Mapping (Grad-CAM) visualizations for defect detection in echocardiograms. Each row shows the comparison of attention maps across five models: the proposed ST-R(2 + 1)D-ConvNeXt and four baseline models (3D ResNet, R(2 + 1)D ResNet, 3D ConvNeXt-V2, and R(2 + 1)D ConvNeXt-V2). (A)-(D): high-quality cases with well-defined chamber structures; (E)-(H): low-quality cases showing various defects, including (E) poorly positioned chambers; (F) distorted contours; and (G)-(H) missing left ventricular walls. Warmer colors (red) indicate regions of higher attention by the model.

proposed ST-R(2 + 1)D-ConvNeXt model. We are the first to intelligently dissect the complexities of spatiotemporal attributes and echocardiographic hallmarks, each contributing distinctively to dynamic echocardiography. Our method surpasses traditional 3D methods,

demonstrating an accuracy of 82.63 % and an AUC of 0.89 in quality classification. Through Grad-CAM visualizations, we provide interpretable feedback on specific quality issues, enabling immediate corrective actions during image acquisition. This systematic approach serves as an early warning system for potentially unreliable measurements while standardizing quality assessment across different operators and institutions. The immediate quality feedback not only streamlines clinical workflows by reducing the need for repeat examinations but also enhances diagnostic confidence through consistent measurement validation. Our work advances the underexplored domain of echocardiography quality surveillance, offering a practical solution for time-sensitive clinical settings where both efficiency and reliability are paramount.

CRedit authorship contribution statement

Chin-Chieh Hsu: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Conceptualization. **You-Wei Wang:** Writing – review & editing, Validation, Supervision, Project administration, Investigation. **Lung-Chun Lin:** Validation, Resources, Funding acquisition, Data curation, Conceptualization. **Ruey-Feng Chang:** Supervision, Resources, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported by National Science and Technology Council of Taiwan (NSTC 113-2221-E-002-167, NSTC 113-2634-F-002-002, and NSTC 110-2314-B-002-210-MY3). C.C. Hsu gratefully thanks the Section of Cardiology, Dept. of Internal Medicine in National Taiwan University Hospital (<https://www.ntuh.gov.tw/>) for their invaluable support in image collection and data interpretation. Additionally, special thanks are extended to the Medical Image Processing Lab (<http://cad.csie.ntu.edu.tw/rfchang/index.php>) in the Dept. of Computer Science and Information Engineering at National Taiwan University (Taipei, Taiwan) for their assistance in providing the necessary computing environment for this research.

Data availability

The data that has been used is confidential.

References

- [1] M.N. Berman, C. Tupper, A. Bhardwaj, Physiology, left ventricular function, StatPearls [Internet] (2019).
- [2] J. Tromp, et al., Automated interpretation of systolic and diastolic function on the echocardiogram: a multicohort study, *The Lancet Digital Health* 4 (1) (2022) e46–e54.
- [3] S. Leclerc, et al., Deep learning for segmentation using an open large-scale dataset in 2D echocardiography, *IEEE Trans. Med. Imaging* 38 (9) (2019) 2198–2210.
- [4] J. Zhang, et al., Fully automated echocardiogram interpretation in clinical practice: feasibility and diagnostic accuracy, *Circulation* 138 (16) (2018) 1623–1635.
- [5] A. Madani, J.R. Ong, A. Tibrewal, M.R. Mofrad, Deep echocardiography: data-efficient supervised and semi-supervised deep learning towards automated diagnosis of cardiac disease, *NPJ Digital Med.* 1 (1) (2018) 1–11.
- [6] K.-C. Huang, et al., Artificial intelligence aids cardiac image quality assessment for improving precision in strain measurements, *Cardiovascular Imaging* 14 (2) (2021) 335–345.
- [7] A.H. Abdi, et al., Quality assessment of echocardiographic cine using recurrent neural networks: feasibility on five standard view planes, in: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11–13, 2017, Proceedings, Part III 20*, Springer, 2017, pp. 302–310.

- [8] Z. Akkus, et al., Artificial intelligence (AI)-empowered echocardiography interpretation: a state-of-the-art review, *J. Clin. Med.* 10 (7) (2021) 1391.
- [9] Z. Liao, et al., On modelling label uncertainty in deep neural networks: Automatic estimation of intra-observer variability in 2d echocardiography quality assessment, *IEEE Trans. Med. Imaging* 39 (6) (2019) 1868–1883.
- [10] L.S. Chow, R. Paramesran, Review of medical image quality assessment, *Biomed. Signal Process. Control* 27 (2016) 145–154.
- [11] G. Valvano, A. Chartsias, A. Leo, and S. A. Tsaftaris, “Temporal consistency objectives regularize the learning of disentangled representations,”. In: *Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data: First MICCAI Workshop, DART 2019, and First International Workshop, MIL3ID 2019, Shenzhen, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13 and 17, 2019, Proceedings 1*, 2019: Springer, pp. 11–19.
- [12] A.H. Abdi, et al., Automatic quality assessment of echocardiograms using convolutional neural networks: feasibility on the apical four-chamber view, *IEEE Trans. Med. Imaging* 36 (6) (2017) 1221–1230.
- [13] C. Luong, et al., Automated estimation of echocardiogram image quality in hospitalized patients, *Int. J. Cardiovasc. Imaging* 37 (2021) 229–239.
- [14] G.E. Jansen, B.D. de Vos, M.A. Molenaar, M.J. Schuurings, B.J. Bouma, I. Išgum, Automated echocardiography view classification and quality assessment with recognition of unknown views, *J. Med. Imaging* 11 (5) (2024) 054002.
- [15] H. Yang, L. Wright, T. Negishi, K. Negishi, J. Liu, T.H. Marwick, Research to practice: assessment of left ventricular global longitudinal strain for surveillance of cancer chemotherapeutic-related cardiac dysfunction, *J. Am. Coll. Cardiol. Img.* 11 (8) (2018) 1196–1201.
- [16] K. Kalam, P. Otahal, T.H. Marwick, Prognostic implications of global LV dysfunction: a systematic review and meta-analysis of global longitudinal strain and ejection fraction, *Heart* 100 (21) (2014) 1673–1680.
- [17] T. Ishizu, et al., Left ventricular strain and transmural distribution of structural remodeling in hypertensive heart disease, *Hypertension* 63 (3) (2014) 500–506.
- [18] R.M. Lang, et al., Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging, *European Heart Journal-Cardiovascular Imaging* 16 (3) (2015) 233–271.
- [19] D.K.S. Rangaraju, K. Kumar, C. Renumadhavi, Review paper on quantitative image quality assessment—medical ultrasound images, *Int. J. Eng. Research and Technology*. 1 (4) (2012).
- [20] V. Mor-Avi, et al., Current and evolving echocardiographic techniques for the quantitative evaluation of cardiac mechanics: ASE/EAE consensus statement on methodology and indications endorsed by the Japanese Society of Echocardiography, *Eur. J. Echocardiogr.* 12 (3) (2011) 167–205.
- [21] N. Painchaud, N. Duchateau, O. Bernard, P.-M. Jodoin, Echocardiography segmentation with enforced temporal consistency, *IEEE Trans. Med. Imaging* 41 (10) (2022) 2867–2878.
- [22] A. Chartsias, et al., Disentangled representation learning in cardiac image analysis, *Med. Image Anal.* 58 (2019) 101535.
- [23] J. Lin, W. Xie, L. Kang, H. Wu, Dynamic-guided spatiotemporal attention for echocardiography video segmentation, *IEEE Trans. Med. Imaging* (2024).
- [24] X. Gao, W. Li, M. Loomes, L. Wang, A fused deep learning architecture for viewpoint classification of echocardiography, *Inf. Fusion* 36 (2017) 103–113.
- [25] S.S. Ahn, et al., Co-attention spatial transformer network for unsupervised motion tracking and cardiac strain analysis in 3D echocardiography, *Med. Image Anal.* 84 (2023) 102711.
- [26] M. Jalali, H. Behnam, F. Davoodi, M. Shojaeifard, Temporal super-resolution of 2D/3D echocardiography using cubic B-spline interpolation, *Biomed. Signal Process. Control* 58 (2020) 101868.
- [27] A. Østvik et al., 2021 “Myocardial function imaging in echocardiography using deep learning,” *IEEE transactions on medical imaging*. 40(5) 1340–1351.
- [28] J. Yang, et al., The efficiency of a Machine learning approach based on spatio-temporal information in the detection of patent foramen ovale from contrast transthoracic echocardiography images: a primary study, *Biomed. Signal Process. Control* 84 (2023) 104813.
- [29] M. Mokhtari, N. Ahmadi, T.S. Tsang, P. Abolmaesumi, R. Liao, Gemtrans: A general, echocardiography-based, multi-level transformer framework for cardiovascular diagnosis, in: *International Workshop on Machine Learning in Medical Imaging*, Springer, 2023, pp. 1–10.
- [30] R.B. Labs, A. Vrettos, J. Loo, M. Zolgharni, Automated assessment of transthoracic echocardiogram image quality using deep neural networks, *Intelligent Medicine* 3 (03) (2023) 191–199.
- [31] R. B. Labs, M. Zolgharni, and J. P. Loo, “Echocardiographic Image Quality Assessment Using Deep Neural Networks,”. In: *Medical Image Understanding and Analysis: 25th Annual Conference, MUA 2021, Oxford, United Kingdom, July 12–14, 2021, Proceedings 25*, 2021: Springer, pp. 488–502.
- [32] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, M. Paluri, A closer look at spatiotemporal convolutions for action recognition, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6450–6459.
- [33] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A convnet for the 2020s, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11976–11986.
- [34] S. Woo, et al., Convnext v2: Co-designing and scaling convnets with masked autoencoders, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 16133–16142.
- [35] K. Kusunose, et al., Reduced variability of visual left ventricular ejection fraction assessment with reference images: the Japanese Association of Young Echocardiography Fellows multicenter study, *J. Cardiol.* 72 (1) (2018) 74–80.
- [36] D. Medvedofsky, et al., Three-dimensional echocardiographic automated quantification of left heart chamber volumes using an adaptive analytics algorithm: feasibility and impact of image quality in nonselected patients, *J. Am. Soc. Echocardiogr.* 30 (9) (2017) 879–885.
- [37] M. Kurt, et al., Impact of contrast echocardiography on evaluation of ventricular function and clinical management in a large prospective cohort, *J. Am. Coll. Cardiol.* 53 (9) (2009) 802–810.
- [38] D. Ouyang, et al., Video-based AI for beat-to-beat assessment of cardiac function, *Nature* 580 (7802) (2020) 252–256.
- [39] D. Ouyang et al., “Echonet-dynamic: a large new cardiac motion video data resource for medical machine learning,”. In: *NeurIPS ML4H Workshop: Vancouver, BC, Canada*, 2019.
- [40] C. Antzelevitch, et al., Heterogeneity within the ventricular wall. Electrophysiology and pharmacology of epicardial, endocardial, and M cells, *Circ. Res.* 69 (6) (1991) 1427–1449.
- [41] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.
- [42] M. De Craene, et al., Temporal diffeomorphic free-form deformation: application to motion and strain estimation from 3D echocardiography, *Med. Image Anal.* 16 (2) (2012) 427–450.
- [43] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization. *arXiv Preprint arXiv:1412.6980*, 2014.
- [44] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
- [45] W. Yu, P. Zhou, S. Yan, X. Wang, Inceptionnext: when inception meets convnext, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 5672–5683.
- [46] Z. Huang et al., “Tada! temporally-adaptive convolutions for video understanding,” *arXiv preprint arXiv:2110.06178*, 2021.
- [47] C. Mitchell, et al., Guidelines for performing a comprehensive transthoracic echocardiographic examination in adults: recommendations from the American Society of Echocardiography, *J. Am. Soc. Echocardiogr.* 32 (1) (2019) 1–64.
- [48] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 618–626.