Real-time Automatic M-mode Echocardiography Measurement with Panel Attention

DOI:
10.1109/JBHI.2024.3413628

Document Version
Accepted author manuscript

Real-time Automatic M-mode Echocardiography Measurement with Panel Attention

Ching-Hsun Tseng, Shao-Ju Chien, Po-Shen Wang, Shin-Jye Lee, Bin Pu, and Xiao-Jun Zeng

***Abstract*— Motion mode (M-mode) echocardiography is essential for measuring cardiac dimension and ejection fraction. However, the current diagnosis is time-consuming and suffers from diagnosis accuracy variance. This work resorts to building an automatic scheme through well-designed and well-trained deep learning to conquer the situation. That is, we proposed RAMEM, an automatic scheme of real-time M-mode echocardiography, which contributes three aspects to address the challenges: 1) provide MEIS, the first dataset of M-mode echocardiograms, to enable consistent results and support developing an automatic scheme; For detecting objects accurately in echocardiograms, it requires big receptive field for covering long-range diastole to systole cycle. However, the limited receptive field in the typical backbone of convolutional neural networks (CNN) and the losing information risk in non-local block (NL) equipped CNN risk the accuracy requirement. Therefore, we 2) propose panel attention embedding with updated UPANets V2, a convolutional backbone network, in a real-time instance segmentation (RIS) scheme for boosting big object detection performance; 3) introduce AMEM, an efficient algorithm of automatic M-mode echocardiography measurement, for automatic diagnosis; The experimental results show that RAMEM surpasses existing RIS schemes (CNNs with NL & Transformers as the backbone) in PASCAL 2012 SBD and human performances in MEIS. The implemented code and dataset are available at https://github.com/hanktseng131415go/RAMEM.**

***Index Terms*— M-mode Echocardiography, Ultrasound Images, Real-time Instance Segmentation**

This work was funded by Kaohsiung Chang Gung Memorial Hospital, Kaohsiung, Taiwan (CORPG8L0101) and Taiwan National Science and Technology Council grant number NSTC 112-2410-H-A49 -033. The study was approved by the Institutional Review Board of Chang Gung Medical Foundation in Taipei, Taiwan (IRB No: 202001552B0C1 and 202101520B0), and all procedures were performed according to the guidelines of the Declaration of Helsinki.

C. H. Tseng is the co-first author for the contribution of the methodology/algorithm development and empirical analysis with the Department of Computer Science, The University of Manchester, Manchester M13 9PR, UK (e-mail: ching-hsun.tseng@postgrad.manchester.ac.uk).

S. J. Chien is the co-first author for the contribution of providing the data and medical domain knowledge with the following affiliations: Division of Pediatric Cardiology, Department of Pediatrics, Kaohsiung Chang Gung Memorial Hospital, Kaohsiung, Taiwan; School of Traditional Chinese Medicine, Chang Gung University College of Medicine, Tao-Yuan, Taiwan; and Department of Early Childhood Care and Education, Cheng Shiu University, Kaohsiung, Taiwan (e-mail: csjdc@cgmh.org.tw).

P. S. Wang and S. J. Lee was and is, respectively, with the Institute of Management of Technology, National Yang Ming Chiao Tung University, Hinschu 30010, Taiwan (e-mail: j2655926@gmail.com; camhero@gmail.com).

B. Pu is with the Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong, China (e-mail: eebinpu@ust.hk).

X. J. Zeng is the corresponding author with the Department of Computer Science, The University of Manchester, Manchester M13 9PR, UK (e-mail: x.zeng@manchester.ac.uk).

I. Introduction

MORE than 90% of heart problems can be detected through the cardiac ultrasound examination [1]. For fetal congenital heart disease, echocardiography is one of the most common methods for diagnosis. Echocardiography includes a 2-D image (B-mode) and a 1-D view of the cardiac motion period (M-mode). Among them, M-mode has been widely used for examining wall thickness, ejection fraction, and other indices based on observing the recording aortic valve (AV) or left ventricle (LV) diastole to systole cycle [2]. M-mode has the advantage of measuring the period by easily marking the peak and the lowest point among a captured M-mode image [3] [4]. It is beneficial to pinpoint the exact cycle beginning and end to get an ideal diagnosis. However, several problems in practical pediatric echocardiography make it nearly impossible, such as the diverse capturing angles, patient resistance, a variance of manual labeling, and time-consuming labeling [5]. Most importantly, one of the notorious issues contributing to the inability of seamless diagnosis is the time-consuming manual labeling [6]. Thus, it brings the following side effects: diagnosis variance and inaccuracy. This work solves the problems, so we propose RAMEM, a Real-time Automatic M-mode Echocardiography Measurement scheme, seeing the overall framework compared with the traditional pipeline in Fig. 1.

This work aims to create an accurate and effective clinical examination by proposing and developing an automatic computer vision technology. To make such automation available, a data-sufficient dataset for training a deep learning workflow should be provided as the foundation for the research. Unfortunately, such a dataset is lacking in the field. Currently, there is a series of automatic deep learning works [7] [8] training on Cardiac Acquisitions for Multi-structure Ultrasound Segmentation dataset (CAMUS) [9], which is a dataset toward B-mode. Still, there is no accessible dataset (work) for M-mode echocardiography to the best of our knowledge. While some works proposed methods for M-mode echocardiography [4] [10], none has made the fully implemented code or dataset publicly available to recreate the results. On the one hand, while most works estimate LV indicators using B-mode spectral flow Doppler based on the predicted segmentation, a direct M-mode examination still remains a precise linear measurement of cardiac dimensions and the most intuitive non-invasive way to assess cardiac function and wall thickness

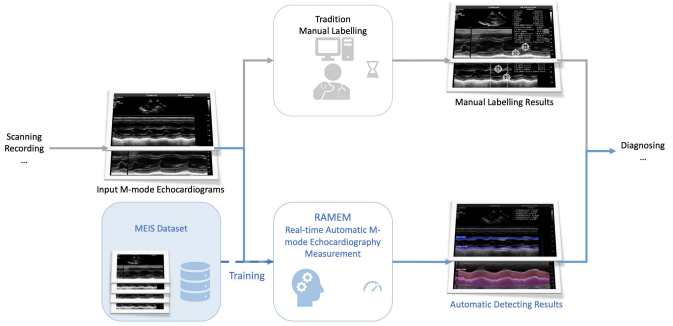

Fig. 1: The proposed framework of RAMEM. The current diagnosis process is in grey, and the proposed one is in blue. The proposed process involves being trained by the proposed MEIS dataset and then detecting the learned labeling agreement from experts to fulfil the automatic echocardiogram measurement in real-time.

[1] [4]. In that context, we managed to collect the daily clinic examination data into the first and biggest M-mode echocardiogram dataset. The dataset makes it possible to ensure consistent diagnosis results and further flourish the following downstream developments.

When it comes to a following development based on a dataset, a natural step is utilizing the existing scheme and optimising it for the targeted field. This work resorts to real-time instance segmentation (RIS) to make a real-time anchoring workflow. For guaranteeing a robust result, a mature scheme is an excellent stepping-stone to building the measurement. In the field, You Only Look At CoefficienTs (YOLACT) [11] [12] is the first and most robust one, which outputs bounding boxes, instant masks, and object labels in real-time under one customer-based GPU. Thus, real-time criteria enable seamless diagnosis; bounding boxes are responsible for locating objects and removing noise; and segmented instant masks contribute to the following indices measurement based on the texture. Despite the merits of YOLACT, how to ensure good performance in such a speedy workflow is still a challenge. Particularly, once a scheme can reach real-time criteria, the importance of Frame Per Second (FPS) (generally >24 FPS as the criteria of being real-time) degrades as the number grows. For example, in daily clinic examination, the experience from a 24 FPS detection is not inferior to the one with 30 FPS, as long as the diagnosis outputs are accurate. In that sense, chasing better detection is the next mission and can make the foundation more solid for the following measurement. For the mission, updating the backbone to an advanced one is helpful, seeing from CenterMaks [13] and Real-Time Models for object Detection (RTMDet) [14]. However, viewing the current YOLACT workflow and the nature of detecting the long-range scope of a cycle in M-mode echocardiography, the backbone suffers from the limited receptive field [15] and could not be the ideal choice for echocardiograms because of the convolutional neural network (CNN)-based ResNets [16]. The limited receptive field could fail to connect essential patterns in the long range, and thus, it could further cause

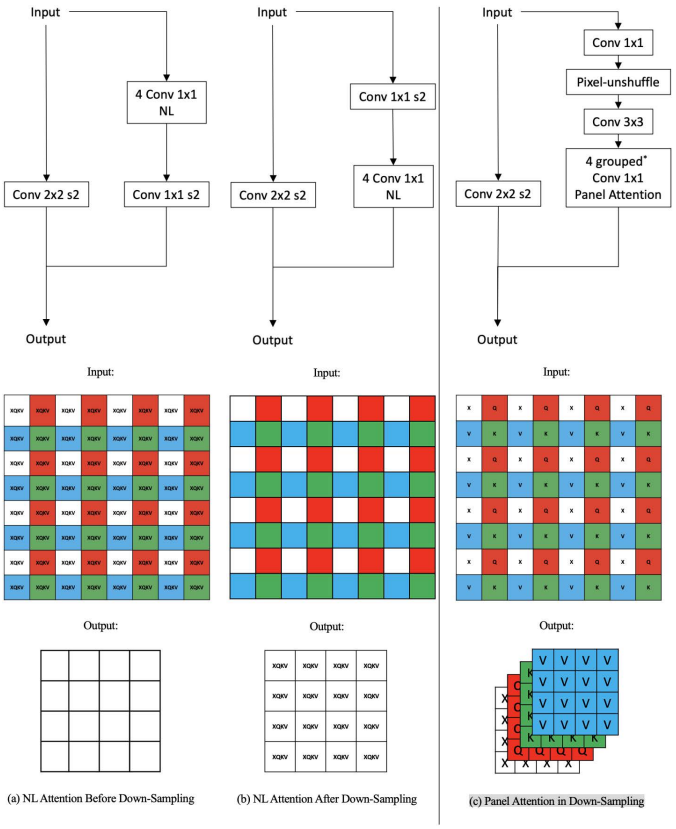

Fig. 2: Panel attention comparison. Different NL attention scenarios in (a) and (b), where (a) operates NL attention on the linear operation (weight assignment) of XQKV firstly and down-sample in 1×1 CNN stride=2 followingly, this makes only white pixels with global information pass through, please see an explanation of XQKV of NL in (1); (b) processes in the opposite way, this makes only white pixels involve global attention; (c) assigns weight by the same linear operation but in a square-clockwise order like panel, which creates a group-like weight connection toward different value (*). QKV is the typical linear operation, and X is the linear operation toward original information or skip-connection value toward output.

misdetection and drag down speed. Deep enough CNNs can indeed have a receptive field covering a whole picture. Still, it will also make a model cumbersome and risk losing information when passing information to deeper layers [16] even with residual connection or dense connection causing performance degradation [17]. While Vision in Transformer (ViT) [18] and non-local block (NL) [15] attention have demonstrated that a series of dot-products across pixels can make global attention, the sacrifice is high computational overhead (attention before down-sampling) that makes the process unable to be real-time. To remedy the inefficiency, ViT makes input images into a much smaller size by a drastic down-sampling (generally a $\frac{1}{16}$ input size in patch=16), and NL applies global attention after down-sampling of each block in default. We argue that such an approach will lose too much information and thus cause a performance downgrade for giving efficient attention when applying to a down-sampled size (attention after down-sampling). In this context, it becomes a dilemma

as having complete non-local without down-sampling will sacrifice efficiency, and having efficient non-local attention will lose information. A question thusly arises: *Is there global attention operating without sacrificing efficiency and losing information?* Our answer is: Panel attention, depth-wise local-to-global attention by pixel-unshuffling. Panel attention creates a third category of *attention in down-sampling*. The intuitive is that, despite the success of Transformer-based models, CNN+NL still outshines in terms of efficiency and trainability in datasets below big scale [19] [20]. As a result, the key is to keep the information passing as much as possible in attention by optimizing NL, which is embedded with a CNN. With the proposed panel attention in our developed UPANets [21], the whole operation is robust and efficient with the recognised merits in both CNN and NL. Please see the demonstration and comparison of this discussion in Fig. 2.

The last puzzle piece is an automatic algorithm for different views of M-mode echocardiograms to shape a real-time automatic M-mode echocardiography scheme. The current complete diagnosis process involves: scanning, recording M-mode video, capturing one clear frame, manually locating anchors, and diagnosing the indicators in the next session. As some diagnoses involve two recording views, the examination should cost double the time. Due to such a process, it is already time-consuming. Even worse, the labeling variance is an inevitable problem across different operators [5] [4]. Automatic deep learning detection and measurement can be the answer for making a seamless diagnosis and maintaining consistency that helps to release the burden from medical professionals. The algorithm to be proposed and developed is called AMEM, automatic M-mode echocardiography measurement, and could create an ideal environment by directly starting a diagnosis from an instant-showing wanted indices based on the current ultrasound image. Merging the algorithm in this work fills the gaps and overcomes the existing issues in M-mode echocardiography. The contributions of this work are listed as follows:

- Open-access MEIS, a dataset of M-mode echocardiograms for instance segmentation. The proposed RIS scheme building can diminish the variance upon the dataset with experts' agreement.
- Provide panel attention, a depth-wise local-to-global efficient attention by pixel-unshuffling, embedding with updated UPANets V2 in RIS with the global receptive field.
- Propose an efficient algorithm of AMEM targeting fast and accurate automatic labeling among diagnoses.

II. Related Work

This section contains the related works from echocardiogram datasets, automatic echocardiography, RIS, and the local & global attention backbone in computer vision. The details can be seen in the following subsections.

A. M-mode Echocardiography Datasets and Automatic Echocardiogram

Echocardiogram Datasets – The direction [22] of automatic echocardiogram detection has recently shrunk into focusing on one direction, B-mode. A survey work [22] on the development of automatic echocardiograms since 2004 has supported this view by showing the distribution: 63 works in B-mode, 28 in Doppler echo, and only 3 in M-mode. Even worse, viewing three datasets for B-mode in 63 works, there is no dataset for M-mode.

Automatic Echocardiography – Applying an automatic scheme to diminish the bias from a human operator is one of the intuitive approaches. Still, the state quo in M-mode diagnoses remains manual operation. Although EchoNet [23] [24] follows a similar notion to eliminate the bias or some works [7] [25] try to estimate the LV mass from B-mode, they are still un-paired with the intuitiveness and accuracy by directly viewing M-mode. This situation could explain why the daily clinic examination of M-mode still stays in manual labeling. To the best of our knowledge, there are only three works that try to break the ice from mode classification, machine learning, and deep learning for animals in order: 1) vanilla fully-connected CNN [26] is applied to classify 15 modes of images (including M-mode) for replacing manual categorizing and speeding up the future down-stream work; 2) The work of using pair-wise distance offsets in [27] might be the first approach try to segment the anterior and posterior wall from a vessel in a motion mode image but the downside of noise disturbing could cloud the performance; 3) Mouse-Echocardiography Neural Net (MENN) [10] by Pfizer faces the automatic issue by proposing the measurement algorithm after deep learning outputs toward animal B-mode and M-mode echocardiography. As a result, MENN might be the closest one to our work. However, the MENN algorithm still involves some manual hyperparameter setting (e.g., setting a sampling period) to detect tissue boundaries. In that sense, whether the setting is suitable for newborns and can be operated in real-time or not remains a question.

B. Real-time Instance Segmentation

Instance segmentation is the task of classifying the pixel category on an image. Real-time generally indicates a whole process from input to output after non-maximum suppression (NMS) is under roughly 0.033 sec (>30 FPS) [11] or 0.042 (>24 FPS) [28]. YOLACT [11], a single-stage anchor-based instance segmentation model, first arrived at the real-time standard. The output includes classes, bounding boxes, and masks. The module for each part is backbone: ResNets50; neck: Feature Pyramid Network (FPN); heads: shallow RetinaNet head; label assignment: anchor-based IoU assignment; NMS: FastNMS. By Adaptive Training Sample Selection (ATSS) [29], if aligning every part of the framework with a dynamic label assignment, the performance has no significant difference from the anchor-based and anchor-free framework. Another YOLACT-based framework with ATSS and the mask-aware intersection of union (maIoU), maYOLACT [28], had become the best performance work in RIS in 2022 and has concreted the standpoints in ATSS. Recently, RTMDet [14] has tried to optimize every submodule based on YOLOX [30], which uses CSPDarkNet, to be the latest model in RIS. Since the introduction of YOLACT, every work ends

with finetuned ResNets from ImageNet pre-trained weight to gain a superior performance. However, as the pre-trained weight from ImageNet significantly differs from the nature of echocardiography, applying such backbones will not benefit our task. Moreover, the same intention could restrain the development of the global attention module in this field.

C. Attention in Computer Vision

Local Attention – Based on the aggregation of the kernel, CNN can be viewed as local attention because the attention area is constrained into k×k in comparing the result of NL. From a broad perspective, it can be said that most CNNs belong to this category. ConvNeXt [31] has argued that modifying ResNets with modern techniques can make CNN outshine Transformer-based networks, such as Swin Transformer [32]. However, from a narrow viewpoint, Local Relation Network (LR-Net) [33] argues that local relations are also vital. It generates appearance composability from the local connection by applying Softmax toward a specific dimension. Similarly, Stand-Alone Self-Attention (SASA) [34] replaced CNNs with the proposed local attention to prove the concept with superiorities.

Global (Non-local) Attention – The introduction of ViT and NL has caused great attention on global attention toward an image. To begin with, capturing such information from a series of dot-products indeed contains spatial and channel information into one. However, we state that ViT (and the variants like Swin Transformers) is different from NL, as ViT turns images into way smaller images as patches with the following a series of multi-head attentions. On the contrary, NL processes the CNN product at every end of each block with single-head attention. Therefore, such a pure attention-based structure of ViT has different features from NL in terms of image size, operating structure, efficiency, and trainability in a small-scale dataset [19] [20]. Therefore, the following discussion mainly focuses on NLs, particularly NLs embedding with CNNs. Taking a 2D image feature $X \in \mathbb{R}^{c \times s}, s = w \times h$ as an example, the NL can be expressed as follows:

$$Y = \left[Softmax\left(Q^T \otimes K,\ s\right)\right] \otimes V^T, \tag{1}$$

where $Softmax(input, dim)$, and $Y, Q, K, V \in \mathbb{R}^{c \times s}$, especially Q, K, V belonging the products of weight assign by linear operations from X, which will be discussed in (2). By (1), the spatial and channel information is aggregated to the product. Nonetheless, because of the notorious high computation overhead, this hardware-unfriendly method is unsuitable for an efficiency-demanded environment. Thus, there can be three categories to deal with global information efficiently: 1) CNN-mimics, 2) Algorithm-simulators, and 3) NL-variants. 1) refers to works claiming that using a big enough kernel can catch the same effect as NL, seeing dilated convolution in Visual Attention Network (VAN) [35] and VAst-receptive-field Pixel attention network (VapSR) [36], but we do not view CNN that merely using a big kernel in this category, e.g., ConvNeXt; The standpoint of 2) is simulated non-local attention can be generated from a delicate algorithm: Expectation-Maximization Attention (EMANet) [37], the authors described multiple iterations of expectation-maximization with hyper-parameters can find the robust attention status, the similar path can be seen in HAMs [38]; 3) is a modification upon NL. Among the modifications, A2 [39] changes the operation of NL in doing $\left[Softmax\left(Q, s\right) \otimes K^T\right] \otimes Softmax(V, c)$ makes the complexity lower. On the one hand, GCNet [40] argues that attention maps across c arguably perform an identical effect, and thus the need to operate in c is unnecessary. Therefore, Sueeze-and-Excitation Network (SENet) [41] following simplified NL in $X' \in \mathbb{R}^{1 \times s}$ should be enough. EANet [42] pushed this direction even further with two layers of perceptron accompanying with $Softmax(X', s), X' \in \mathbb{R}^{\frac{c}{64} \times s}$ as inter feature maps. We put our panel attention in category 3).

TABLE I:
Data distribution of MEIS.

Type	Object (mask)	Indicators	Training/Testing (number)
Aortic Valve (AV)	Aortic Root (AoR), Left Atrium (LA)	AoR Diameter, LA Dimension	747/559
Left Ventricle (LV)	Interventricular Septum (IVS), Left Ventricular Posterior Wall (LVPW)	Left Ventricular Internal Diameter end systole (LVIDs), Left Ventricular Posterior Wall end systole (LVPWs), Interventricular Septal end systole (IVSs), Left Ventricular Internal Diameter end diastole (LVIDd), Left Ventricular Posterior Wall end diastole (LVPWd), Interventricular Septal end diastole (IVSd)	774/559

III. MEIS Dataset

This work presents a dataset for bridging M-mode echocardiography and RIS. To our knowledge, it is the first M-mode echocardiogram dataset. Therefore, we name this dataset MEIS, M-mode echocardiograms for instance segmentation. Table I and the following sub-sections reveal the dataset distribution and details.

A. Description

MEIS is an echocardiogram dataset that comprises a total of 2,639 images from a total of 923 de-identified subjects in the image size of 1024×768 toward two recording views (AV and LV) with 1,521 (747 in AV + 774 in LV) images for training and 1,118 (559 in AV + 559 in LV) for testing, respectively. Each view must be detected with two objects to calculate the measurement indicators. That is in total with four object classes (two objects in each view): aortic root (AoR) and left atrium (LA) in AV; interventricular septum (IVS) and left ventricular posterior wall (LVPW) in LV. The medical meaning and purpose of each indicator are listed in the following:

- AV: LA-Dimension and AoR-Dimension can be measured for calculating different indicators, such as AoR/LA ratio, to examine the state of the aortic valve.
- LV: 6 measurements include IVSs, IVSd, LVIDs, LVIDd, LVPWs, and LVPWd. These concerned thicknesses and dimensions in LV recording are used to estimate other cardiac functions through specific medical formulas, including LV mass, LV ejection fraction, end-diastolic volume, end-systolic volume, and more [43] [44].

B. Preparation

The source of images is collected and approved within the regulation set by the ethical committee of the Chang

Gung Medical Foundation Institutional Review Board. The permission to open the dataset to the public and de-identified patients' information has also been approved. The dataset is recorded from Philips EPIQ7 and iE33 ultrasound machines. It comprises daily clinic images of children under 18, including infants. Our objective is to facilitate the development of automated measurement with the proposed scheme. Hence, the dataset encompasses images of our pediatric patients without specific selections, which will not compromise our essential care in measurement and subsequent analysis based on M-mode echocardiography. The de-identified process involves removing patient details and replacing each data name with the dataset-specific file name that cannot be tracked back to the original patient. However, as there are cases of one patient with two examination views, these will be assigned with the same file name with different sub-numbers. After de-identification under three physicians' supervision and cross-validating the results, the preparation results require object locations, object classes, and object masks. The data preparation, thus, is done with physicians through a series of preparations, from manually drawing masks to file conversion under the same supervision and validation.

Manual Mask Drawing – Masks of objects are critical for image segmentation tasks. For M-mode echocardiograms, the traditional way is to manually set multiple anchors in interested locations among the heart's inner surface or the vessel wall. This method, however, could lead to a wrong location. To avoid this, a careful mask drawing on a still recorded image could be a solution because the mask can represent the accurate belonging boundary from object to object. As the process is locating the marks on one captured image, the video frames, the captured frame is taken as the training image. Thus, in raw AV-recording images, the contours of AoR and LA are drawn manually along with the experts and based on the marks on the original extracted frame. The marks were made during daily clinic examinations. In raw LV-recording images, IVS and LVPW are processed in the same way.

File Conversion (Bounding Box to COCO Format) – Bridging between the M-mode echocardiography and RIS is one of the purposes of this work. Thus, a dataset should follow the mainstream RIS, COCO format in JSON. This action also brings another advantage to M-mode echocardiography: "bounding box". Considering different viewing angles could cause different mask scales, fixed and manual-based postprocessing could cloud the meaning of automation. In contrast, the bounding box of an object could be the perfect solution, as the bounding box already has the merit of focusing on the mask in the bounding box. In this case, the only work left for automation is perfecting the generation of masks. Moreover, the current detection strategy of RIS is also based on a bounding box, making the bounding box even more vital. After the mask drawing, bounding boxes of objects are decided upon the location of the masks. The classes of masks, along with the bounding box location, are also organized in JSON format. Finally, because the images in COCO format are typical in JPG format [45], we transfer the original file of Digital Imaging and Communications in Medicine (DICOM) into JPG. In sum, a COCO format M-mode echocardiography is followed with JPG images, masks, classes, bound boxes, and actual measurement index values by the daily clinic examination in manual anchoring from the above ultrasound machines.

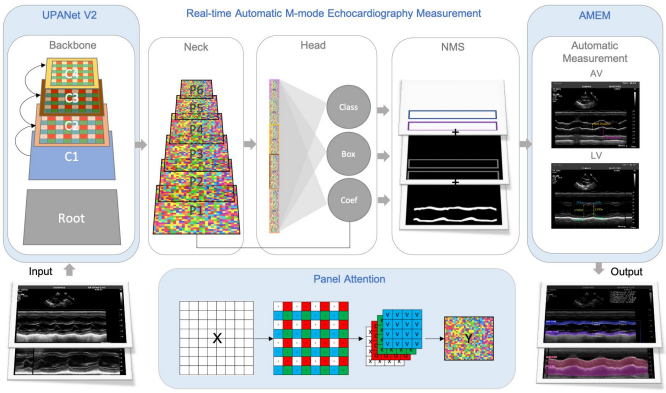

Fig. 3: The Detailed of RAMEM. The proposed methods are in blue. UPANets V2 has been updated with panel attention and equipped with global attention ability. After going through the process from backbone to NMS, the proposed automatic algorithm, AMEM, is for measuring wanted indicators toward M-mode echocardiograms.

IV. RAMEM

The proposed methods in this paper include new local-to-global attention with an updated backbone and an automatic measuring algorithm for M-mode echocardiography. As an overall pipeline compared with the traditional one has been shown in Fig. 1, a more detailed framework demonstration that covers the backbone, attention, and algorithm can be seen in Fig. 3 and the following subsections.

A. Panel Attention

Typically, objects in echocardiograms occupy a great portion of the pixels. However, this scene is not a privilege that big objects also appear in the real-world scene. In fact, having a big receptive field is one of the ultimate goals of computer vision. To showcase our answer, panel attention formulation should be a better candidate in these aspects: 1) lossless operation in pixel-unshuffle, 2) caring local information in CNN, and 3) capturing global information more efficiently. Because pixel-(un)shuffle is one of the famous methods to preserve information while changing the feature map size, a stride=2 ($p = 4$) pixel-unshuffle can make attention in down-sampling. Moreover, the output of NL is a sort of one combination value from four weighted pixels (X, Q, K, V), so making the replacing four pixels from pixel-unshuffle into the four weighted pixels can achieve the issue we want to resolve. As a result, the differences that help our panel attention to outshine other NLs are the 1) lossless operation in pixel-unshuffle. 1) makes attention in down-sampling, so attention can save the burden of facing high spatial computation when attention before down-sampling and chasing better performance without losing information when attention after sampling. Another difference is using depth-wise Conv to implement weight assignment and local attention at the same time, we separate

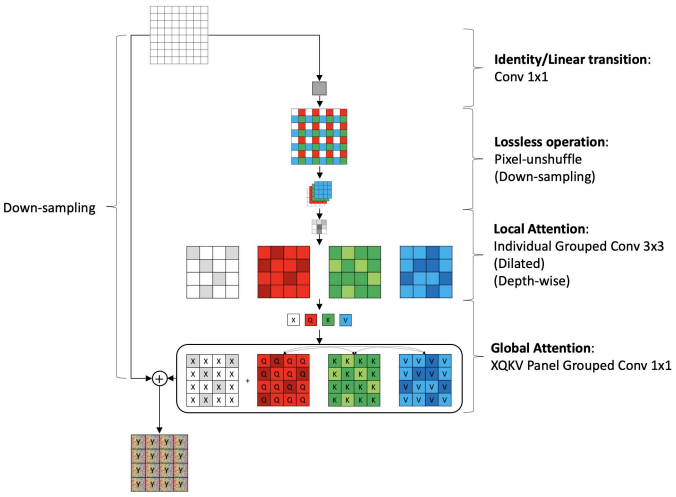

Fig. 4: The components cooperate as a whole in panel attention. Taking the input feature map in $1 \times 8 \times 8$, the output is $1 \times 4 \times 4$. The Conv stride in 1 is ignored for easy presentation.

our depth-wise Conv into the local attention (with grouping Conv 3×3 in 2)) and the global attention (with fully-connected Conv 1×1 in 3)), though. It successfully bridges two parties (depth-wise Conv and NL) and makes two parties compensate each other. Therefore, our attention follows and contains four components from identity/linear transition (Conv 1×1), pixel-unshuffle, local attention (grouped Conv 3×3), and global attention (panel grouped Conv 1×1). Apart from the first transition, the rest of the components shape the characters of the panel attention. Please see the detailed demonstration in Fig. 4 along with the formulation, which is later embedded in updated UPANets to replace the channel and spatial attention as UPANets V2 acting as a backbone of RIS.

Identity/Linear transition – This transition aims to align channel numbers and weight assignment. If the input feature map has the same channel number for the following operation without weight assignment, an identity operation is implemented, and vice versa. The channel alignment and pixel weight assignment are fulfilled by a fully-connected layer. Please see the equation below:

$$X_{_l} = \begin{cases} W\left(X\right), when\, alignment\, and\, assignment \\ X, else \end{cases}, \quad (2)$$

where W refers to linear transition. $X_l \in \mathbb{R}^{c \times s}$ and $X \in \mathbb{R}^{c^{'} \times s}$ when $c^{'}$ is aligned and assigned with weight to c. Otherwise, $X, X_l \in \mathbb{R}^{c \times s}$.

Lossless operation – Pixel-(un)shuffle is known as preserving information among moving spatial and depth bilaterally. This transition by pixel-unshuffle is suitable for the most severe computation overhead issue from the spatial dimension in NL because the spatial information is transferred into the depth. This operation down-samples feature maps in a lossless way as follows:

$$X_s = P^{-}\left(X_l\right), \quad (3)$$

where $X_s \in \mathbb{R}^{4c \times \frac{s}{4}}$, and P^{-} represents a pixel-unshuffle operation. Pixel-unshuffle can be seen in Fig. 4 where different colour pixels in the dilated arrangement are then organized into a small matrix in the belonging colour. However, this transition brings another issue: increasing the channel number (depth) by four times. The remedy for this side effect is presented in global attention by panel grouped Conv 1×1.

Local attention – Having 3×3 Conv upon the unshuffled pixel creates a dilated effect, expanding the receptive field to 5×5; please see the effect in Fig. 4 where pixels separate according to colour. Also, using individual grouped CNN further forms a depth-wise convolution [46] for the following layers. Besides, because of the in-existing normalization and activation between depth-wise convolutional layers for the successive layers, it preserves the local aggregation information to build appearance composability:

$$X_{la} = \theta\left(X_s,\ K\right), \quad (4)$$

where θ refers to the general CNN aggregation with kernel $K \in \mathbb{R}^{c \times k \times k}$, la belonging to the abbreviation of local attention. Default $k = 3$.

Global attention – Continuing the former compounds, two missions must be accomplished as fine attention: remedy the increased channel dimension and make non-local attention. A special grouped weight assign policy, panel group, is proposed to remedy the expanded channel. Unlike generally grouped CNN, the panel group selects each channel in each p step as a group. In other words, the sum of weights equals the general grouped CNN number as 4. Please see the equations:

$$\left\{X_{la}{}', Q_{la},\ K_{la}, V_{la}\ \right\} = W \times select\left(X_{la,i},\ p\right), \quad (5)$$

$$X_{ga} = f\left(Q_{la},\ K_{la}, V_{la}\right), \quad (6)$$

$$Y = \sigma\left(N\left(X_{la}{}' + X_{ga}\right)\right) + X, \quad (7)$$

here $p = 4$ in default, $\left\{X_{la}{}', Q_{la},\ K_{la}, V_{la}\ \right\} \in \mathbb{R}^{4c \times \frac{s}{4}}$ with the same meaning as (1) but in a small size, $X_{la}{}' = WX_{la,\ i=1,5,\cdots n-4}$ as a linear transition. $\{Q_{la},\ K_{la},\ V_{la}\}$ follow the same weight assignment (linear transition) policy as $X_{la}{}'$ but with $i = 2, 3, 4$ as the start, respectively. For example, $X_{la}{}'$ takes the $1st, 5th, \ldots, n-4th$ channels, Q_{la} takes $2nd, 6th, \ldots, n-3th$ channels, etc. f represents the NL block at A2 [39], $\left[Softmax\left(Q, s\right) \otimes K^{T}\right] \otimes Softmax(V, c)$. N and σ refer to normalization and activation separately in skip-connection. ga belongs to the abbreviation of global attention. Finally, the output will be added upon X as the final output.

B. UPANets V2

Considering a more robust performance and the potential inefficiency in ResNet, we opt for an efficiency backbone, UPANets, as the backbone and update upon it in the YOLACT scheme. In ConvNeXt, there are debates about whether to use normalization/activation, the number of times they are used among a block, and how to properly down-sample. The updated propositions, except updating channel & spatial pixel attention by panel attention, are as follows:

- Activation: apply PReLU viewing the parametric can be identity output as $p = 1$ and $0 \leq p < 1$ as typical activation.

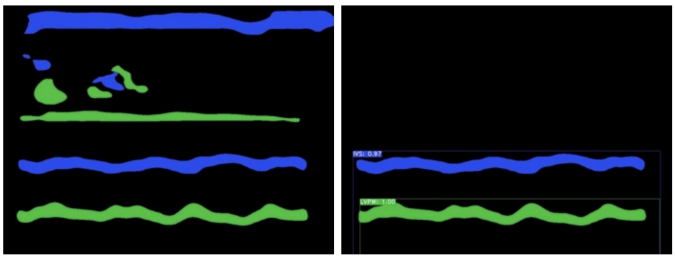

Fig. 5: Postprocessing with a bounding box, the pre-image on the left and the post-image with filtering out un-wanted masks on the right.

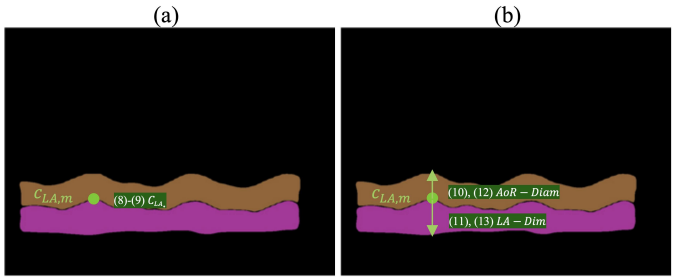

Fig. 6: AV indicators measurement. (a) is the coordinate of the topmost. (b) is the result of finding by scanning up and down through AoR and LA.

- Normalization: except $Softmax$ as global attention normalization, layer normalization is used as spatial normalization and batch normalization is used afterwards to form a double normalization inspired by EANet double normalization policy.
- Down-sampling: a patch (separate) down-sampling of stride 2 Conv 2×2 is applied as skip-connection in origin.

C. Automatic M-mode Echocardiography Measurement

After getting classes, bounding boxes, and masks from the scheme in YOLACT, a post-process is implemented after NMS to eliminate noise outside the bounding boxes. The bounding boxes aim to make the following measurement focus on the wanted area, seeing Fig. 5. Apart from excluding the noise, we opt to detect the whole picture instead of the only echocardiogram at the bottom because of the needed scale (the pixel-to-cm ratio) for the following deducing actual length. Moreover, although it makes more sense to only feed the needed area for detecting, implementing cropping during frame capturing hinders the speed and causes the potential of having an inflexible workflow. Therefore, the desired indicators in Table I with the proposed measuring method are divided into two views and having the following discussion:

- AV: for AoR Diam from AoR; and LA Dim from LA.
- LV: for LVPWs and LVPWd from LVPW (systole/diastole); for IVSs and IVSd from IVS (systole/diastole); for LVIDs and LVIDd among LVPW to IVS (systole/diastole).

Aortic Valve – The AoR Diam and LA Dim indicators serve to observe the state of AV. The character of parallel walls in the aorta moves anteriorly in systole and posteriorly in diastole, so the dimension would not change in either state. The indicators, therefore, can be measured at any place of the detected aortic root mask. Conversely, LA-Dim is generally measured at end-systole, where the LA mask volume is maximum at each period. To do that, this work directly captures the topmost point among the mask. This method involves finding the contour coordinates by topological algorithm [47]. A collective coordinate from FindCountours can be represented as:

$$c_{LA,m} = (x_{LA,m}, y_{LA,m}) = FinedCounters\left(Mask_{LA}\right), \tag{8}$$

$$C_{LA_*} = \max_{-x,y} c_{LA,m}, \tag{9}$$

which $(x_{LA,m}, y_{LA,m})$ stands a mask contour coordinate C in the x-axis and y-axis, and $\max_{-x,y} c_{LA,m}$ is the operation for finding the topmost point by taking the maximum of converted coordinates in $(-x, y)$. Thus, the topmost one should be the first point in the clockwise order. Then, by finding C_{LA_*} in $Mask_{LA}$, going upwards and downwards pixel-wise obtains the total number of vertical pixels of AoR-Diam and LA-Dim as (10) and (11). Finally, multiply these numbers with the echocardiogram image $Scale$, in (12) and (13), to get the actual length. The $Scale$ represents the ratio between actual cm compared with the number of pixels in an image's height, pixel-to-cm ratio. Please see the demonstration from (a) to (b) in Fig. 6 and the equations below.

$$C_{AoR} = Sign\left(Mask_{AoR}\left(C_{LA_*}\right)\uparrow\right), \tag{10}$$

$$C_{LA} = Sign\left(Mask_{AoR}\left(C_{LA_*}\right)\downarrow\right), \tag{11}$$

$$AoR - Diam = Scale \times \left(C_{AoR}\left(y\right) - C_{LA_*}\left(y\right)\right), \tag{12}$$

$$LA - Dim = Scale \times \left(C_{LA_*}\left(y\right) - C_{LA}\left(y\right)\right). \tag{13}$$

Left Ventricle – Examining indicators of LV involves carefully locating where systole and diastole happen. Systole typically indicates the smallest volume, and diastole is the opposite. The same AV strategy can easily gain systole moment to get the most points in LVPW (14) - (15). By scanning up and down, we can sequentially extract LVIDs, IVSs, and LVPWs pixels multiplying with $Scale$, as the following (16) - (21):

$$\begin{aligned} c_{LVPW,m} &= (x_{LVPW,m}, y_{LVPW,m}) \\ &= FinedCounters\left(Mask_{LVPW}\right), \end{aligned} \tag{14}$$

$$C_{LVPW_\sigma} = \max_{-x,y} c_{LVPW,m}, \tag{15}$$

$$C_{LVIDs} = Sign\left(Mask_{background}\left(C_{LVPW_\sigma}\right)\uparrow\right), \tag{16}$$

$$C_{IVSs} = Sign\left(Mask_{IVS}\left(C_{LVIDs}\right)\uparrow\right), \tag{17}$$

$$C_{LVPWs} = Sign\left(Mask_{LVPW}\left(C_{LVPW_\sigma}\right)\downarrow\right), \tag{18}$$

$$LVIDs = Scale \times \left(C_{LVIDs}\left(y\right) - C_{LVPW_\sigma}\left(y\right)\right), \tag{19}$$

$$IVSs = Scale \times \left(C_{IVSs}\left(y\right) - C_{LVIDs}\left(y\right)\right), \tag{20}$$

$$LVPWs = Scale \times \left(C_{LVPW_\sigma}\left(y\right) - C_{LVPWs}\left(y\right)\right). \tag{21}$$

Diastole moment, however, needs some tricky methods to finish. Firstly, find the diastole point, which creates the biggest volume lying in one of the defect points among a convex hull, so structuring the convex hull connection among the mask by Sklansky's algorithm [48] is applied.

$$\begin{aligned} c^H_{LVPW,m} &= \left(x^H_{LVPW,m}, y^H_{LVPW,m}\right) \\ &= ConvexHull\left(Mask_{LVPW}\right). \end{aligned} \tag{22}$$

With the contour points and hulls determined upon $c_{LVPW,m}$, Secondly, defect points among hulls can be extracted by calculating the maximum distance in each hull, (23). Followingly, only picking the upper defect points with the biggest distance among all the defect points as the diastole point is applied, (24). The operation can also contribute to saving time by ignoring the unwanted area in the bottom part of the mask.

$$\begin{aligned} Defects &= \left(starts_{x,y}, ends_{x,y}, defects_{x,y}, distances\right) \\ &= FindDefects\left(c_{LVPW,m}\ ,\ c^H_{LVPW,m}\right). \end{aligned} \tag{23}$$

$$C_{LVPW_\delta} = \max_{sign\left(Defects(ends_x - starts_x)\right)\in +} Defects\left(distances\right). \tag{24}$$

Finally, following the same strategy in scanning up and down pixels as IVSs, LVIDs, and LVPWs, the equations for getting the diastole points of IVSd, LVIDd, and LVPWd along with multiplying $Scale$ are as below:

$$C_{LVIDd} = Sign\left(Mask_{background}\left(C_{LVPW_\delta}\right)\uparrow\right), \tag{25}$$

$$C_{IVSd} = Sign\left(Mask_{IVS}\left(C_{LVIDd}\right)\uparrow\right), \tag{26}$$

$$C_{LVPWd} = Sign\left(Mask_{LVPW}\left(C_{LVPW_\delta}\right)\downarrow\right), \tag{27}$$

$$LVIDd = Scale \times \left(C_{LVIDd}\left(y\right) - C_{LVPW_\delta}\left(y\right)\right), \tag{28}$$

$$IVSd = Scale \times \left(C_{IVSd}\left(y\right) - C_{LVIDd}\left(y\right)\right), \tag{29}$$

$$LVPWd = Scale \times \left(C_{LVPW_\delta}\left(y\right) - C_{LVPWd}\left(y\right)\right). \tag{30}$$

With the well-explained procedure, the demonstration of LV can be seen in Fig. 7. The whole process is organized as the algorithm of AMEM.

V. Experiment

A series of evaluations are conducted to prove the capabilities of the proposed methods, from ablation studies toward panel attention to the real-world indicator bias evaluation based on the provided dataset. Especially, the COCO matric in mean average precision (mAP) is a more complex and thorough benchmark, as each 5-step threshold represents a hesitate (or confidence) index toward an object, compared with naïve indices in the mean of intersection over union (IoU) and DICE (an IoU variant). Because the evaluation task involves not only bounding boxes but also semantic masks, the mAP covers box and mask IoU. Lastly, we opt for >24 FPS (<0.042 sec) as the real-time standard. The details of the specific experiment settings and the discussion of our proposed methods are shown in each sub-section below.

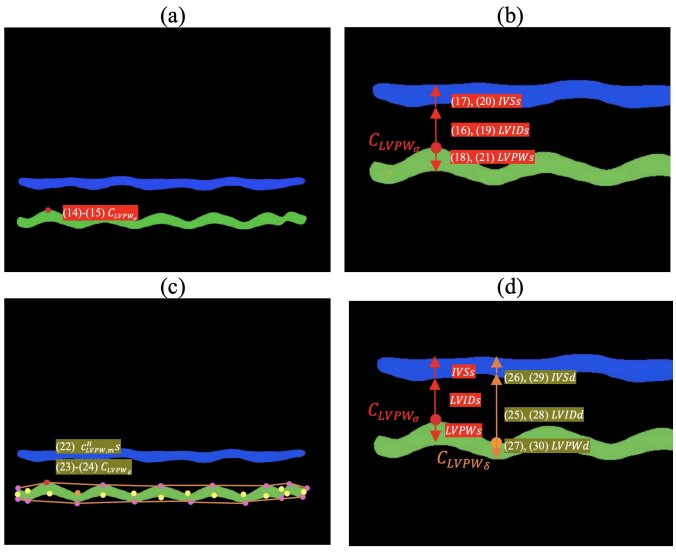

Fig. 7: LV indicators measurement. (a) finds the topmost coordinate in LVPW; (b) searches up and down to get LVIDs, IVSs, and LVPWs; (c) constructs the convex hull with peak points in purple, defects in yellow, and connection line in light orange. Then, pick the one with the largest distance from the defect to the connection line in orange. (d) follows the same strategy to search up and down, based on the finding at (c), and gets LVIDd, IVSd, and LVPWd.

Algorithm AMEM

x: input image
$M \leftarrow f(x)$ real-time model outputs, which contain classes, bounding boxes, and masks.

for m **in** M:
  | **in** view AV:
  || $c_{LA,m}$ = find contour points of $Mask_{LA}$: (8)
  || C_{LA_*} = locate the top most point: (9)
  || $AoR - Diam, LA - Dim$ = search up and down to get indicators based on the scale: (10)-(13)
  || **return** $AoR - Diam, LA - Dim$
  | **in** view LV:
  || $c_{LVPW,m}$ = find contour points of $Mask_{LVPW}$: (14)
  || **in** systole:
  ||| C_{LVPW_σ} = locate the topmost point of $Mask_{LA}$: (15)
  ||| $LVIDs, IVSs, LVPWs$ = search up and down to get indicators based on the scale: (16)-(21)
  || **in** diastole:
  ||| $c^H_{LVPW,m}$ = construct convex hull: (22)
  ||| C_{LVPW_δ} = find the diastole point from defects: (23)-(24)
  ||| $LVIDd, IVSd, LVPWd$ = search up and down to get indicators based on the scale: (25)-(30)
  ||| **return** $LVIDs, IVSs, LVPWs, LVIDd, IVSd, LVPWd$

A. PASCAL 2012 SBD

To evaluate the effect of panel attention, the testing experiment of the existing backbones is conducted first on PASCAL 2012 SBD, a standard open dataset for assessing an object detection model. As this work focuses on real-time instance segmentation, the simulation in a mature scheme, YOLACT, is

TABLE II:
PASCAL 2012 SBD results.

#	Work	Backbone	Avg-mAP↑	Mask-mAP↑	Box-mAP↑	FPS (RTX 4090)↑	Size (M)↓
1*	YOLACT	ResNet101	29.96	29.24	30.67	81.00	51.58
2	maYOLACT	ResNet101	32.16	31.57	32.74	**81.43**	51.45
			+7.34%	+7.97%	+6.75%	+0.53%	-0.25%
3	maYOLACT	CSP-DarkNet	40.29	39.75	40.83	80.36	45.44
			+34.48%	+35.94%	+33.13%	-0.79%	-11.90%
4	maYOLACT	ViT	16.67	16.60	16.73	66.43	93.88
		(Base, batch=16)	-44.36%	-44.59%	-44.16%	-17.99%	+213.35%
5	maYOLACT	Swin Transformer	19.73	19.55	19.90	54.91	94.21
		(Base, batch=4)	-34.15%	-33.14%	-35.12%	-32.21%	+82.65%
6	RTMDet	CSP-DarkNet	32.39	30.56	34.22	68.42	83.33
	(ins-X)[1]		+8.11%	+4.51%	+11.57%	-15.53%	+61.55%
7	RAMEM	UPANet80 V2	41.72	41.51	41.93	64.50	**35.87**
		(Vanilla)	+39.25%	+41.96%	+36.71%	-20.37%	-30.46%
8	RAMEM	UPANet80 V2	42.24	42.20	42.47	63.91	39.30
		(GC, Att 1st)[2]	+40.99%	+44.32%	+38.47%	-21.10%	-23.81%
9	RAMEM	UPANet80 V2	41.86	41.65	42.07	64.50	39.30
		(EA, Att 1st)	+39.72%	+42.44%	+37.17%	-20.37%	-23.81%
10	RAMEM	UPANet80 V2	41.26	41.7	42.62	53.25	40.50
		(A2, Att 1st)	+37.72%	+42.61%	+38.96%	-34.26%	-21.48%
11	RAMEM	UPANet80 V2	-	-	-	-	40.77
		(NL, Att 1st)	-	-	-	-	-20.96%
12	RAMEM	UPANet80 V2	42.35	42.20	42.51	62.86	39.30
		(GC, Ds 1st)[3]	+41.36%	+44.32%	+38.60%	-22.40%	-23.81%
13	RAMEM	UPANet80 V2	41.33	41.04	41.62	64.28	39.30
		(EA, Ds 1st)	+37.95%	+40.36%	+35.70%	-20.64%	-23.81%
14	RAMEM	UPANet80 V2	41.13	40.92	41.33	60.34	40.50
		(A2, Ds 1st)	+27.89%	+29.62%	+26.24%	-25.90%	-21.28%
15	RAMEM	UPANet80 V2	42.24	42.17	42.30	61.64	40.23
		(NL, Ds 1st)	+40.99%	+44.22%	+37.92%	-23.90%	-22.00%
16	YOLACT	ResNet101	38.19	37.57	38.81	68.14	93.58
		(Panel attention)	+27.47%	+28.49%	+26.54%	-15.88%	+81.43%
17	maYOLACT	ResNet101	40.41	40.00	40.81	57.66	93.58
		(Panel attention)	+34.88%	+36.80%	+33.06%	-28.81%	+81.43%
18	RAMEM	UPANet80 V2	**42.69**	**42.42**	**42.96**	60.93	40.32
		(Panel attention)	+42.49%	+45.08%	+40.07%	-24.78%	-21.83%

*The baseline
[1]RTMDet-ins X: Considering the main contributions of RTMDet laying in backbone, neck, and label assignment, we only apply CSPDarkNet X + PAFPN + soft SimOTA. For a fair comparison, the rest of the modules remain as YOLACTs.
[2]Att 1st: Attention before down-sampling
[3]Ds 1st: Attention after down-sampling

TABLE III:
Attention ablation study.

#	Backbone UPANet80 V2 (Without attention)	Local Attention	Global Attention	Panel Attention	Avg-mAP↑
1	-	-	-	-	37.39
2	✓	-	-	-	41.50
3	✓	✓	-	-	41.85
4	✓	-	✓	-	42.51
5	✓	✓	✓	✓ (global-local, p=4)	42.14
6	✓	✓	✓	✓ (local-global, p=1)	41.63
7	✓	✓	✓	✓ (local-global, p=4)	**42.69**

[1]Panel attention: local attention + global attention, both local and global attention having lossless operation involving as attention in down-sampling, please refer Section IV-A

followed. The end-to-end training is set to 64k iterations with the same learning rate, data augmentation, and learning rate schedule in YOLACT. In Table II, The complete maYOLACT and RTMDet are also tested in this simulation, with input shape in $3 \times 544 \times 544$. Note that the model size indicates different backbones, which will affect the FPN input channel number accordingly. The results of mask and box only reported mAP considering the aesthetic, added another index of average mAP from mask and box in Avg-mAP, accompanied with the improvement ratio with the baseline under each value (green for upgrade; red for downgrade), and the general priority of index in order is: Avg-mAP, Mask-mAP, Box-mAP, FPS, and then size.

Vanilla UPANets V2 outperforms other backbones in mAPs (Table II #1 - #7) – This group contains two types of backbone: pure CNNs (ResNets, CSP-DarkNets, and UPANet) and pure NLs (ViT and Swin Transformer). Taking ResNet101 YOLACT #1 as the baseline, maYOLACT #2 substantially boosts performance under the same parameters. CSP-DarkNet in #6 does not experience the same boost, while the result of #3 is promising. What makes the observation surprising is the benefit of the global receptive field of ViT #4 and Swin Transformer #5 is not revealed in this simulation. The possible reason could be responsible for the argued issue of losing information in section I (attention after down-sampling) and the nature of averaging feature maps in Transformers NL blocks [19]. Moreover, the under-scale datasets like PASCAL and medical datasets could be insufficient to ignite the performance of Transformers [49], which could explain the scene and not be suitable for our application. Conversely, by #7, vanilla UPANet80 V2 shares the advantage of trainability of CNNs in this scale of a dataset, having better average mAP and fewer parameters compared with the existing backbones. Therefore, the following comparison will focus on the NL variants in different scenes upon UPANets V2 viewing a better capability of UPANets V2 than other backbones.

Att 1st makes NL inefficiency despite mAPs improvement (Table II #8 - #11) – This group follows the workflow as (a) in Fig. 2, using different NL variants in stage 2, 3, 4 as general. The candidates across GCNet (GC), EANet (EA), A2, and NL use the default setting. #8 in GC has outshined others with the least size, speedy FPS, and mAPs, which aligns with the GCNet results [40]. However, because of the unbearable complexity caused by spatial, which aligns with our statement in section I (attention before down-sampling), #11 in NL was inoperable in a 24G GPU. That issue is bypassed by A2 #10 in operating toward a different dimension order: the channel and then spatial, but the performance is compromised.

Ds 1st makes NL efficiency with experiencing the risk of mAPs degradation (Table II #12 - #15) – Following the workflow as (b) in Fig. 2. This group has experienced a similar trend as the last group (Att 1st) but faces the issue in I (attention after down-sampling), seeing #6. Also, NL #15 outperforms EA #13 and A2 #14, which indicates that using simplified NL might lose some vital features to supplement complexity.

CNNs + Panel attention owns better balance between mAPs and FPS (Table II #16 - #18) – With panel attention equipped in ResNets and UPANet, it brings more benefits (#1 v.s. #16 and #2 v.s. #17) and outperforms existing NLs in either mode (#8 - #15). As our panel attention is in the category of attention in down-sampling, #18 has less FPS degradation than #10 in the category of Att 1st. Ass 1st can be viewed as placing priority on detection accuracy. Although simplified NL-variants of GC and EA have better FPS, their mAPs are dimmed by the performance of panel attention in #18 because of purposely avoiding spatial calculation, causing efficiency degradation in simplified NL-variants. In Ds 1st, NLs can save some computational cost, but they also lose detection performance, so they have worse mAP despite owning better FPS in #12 - #15. Combining both observations and looking into panel attention in #18, it brings the performance even more viewing average, mask, and box map and does not lose too much efficiency in FPS, so we proposed the attention as such.

Panel attention is more general with detailed ablation study (Table III #1 - #7) - Diving into a deeper discussion, a detailed performance contribution can be seen in the ablation study in Table III, which includes: the local attention #3, the global attention #4, and #5 to #7 from the local + global attention with $p = 1$ in (5); flipped attention in global to local; to panel attention as our final attention (local + global attention). Table III #1 aligns to Table II #1 and Table III #2 to Table II #7. The local attention can contribute to mAPs

because the dilated depth-wise CNN expands the receptive field in a dilated effect. The global attention contributes more, thus solidifying that a global receptive field is crucial. Until this observation, the local and global attention can contribute to performance. By merging them together, we can see a better performance, but the improvement is relatively minor. The limited margin of 0.18 between #7 and #3 can be explained by some proportion overlap effect between local and global attention. The local attention uses convolutional layers upon pixel-unshuffle, creating a bigger receptive field. The bigger the receptive field forms in CNN, the smaller the gap between local and global attention receives. Despite that, both sub-attention share improvement and merging together still benefits more, indicating they compensate each other. To view whether the compensation is from the attention order, another observation of flipping attention order in #5 shows degradation. The reason for the decreasing result could be that the following local attention clouds the global attention information. In other words, although the local information contains crucial patterns helping detection, implementing it first negatively affects the global relation information in the feature map, which could explain why most of the NLs place attention after each convolutional block, like (b) in Fig. 2. Finally, a variation test of seeing the source of compensation shows $p = 1$ in (5) #6, which essentially degrades our panel attention into a typical NL [15]. The default setting $p = 4$ ought to select the corresponding pixel because of pixel-unshuffle (by expanding the channel in advance in 3) and make sure that the product of the attention is with the same channel number in a down-sampling size. Therefore, $p = 1$ equals to do pixel-unshuffle without channel expanding in advance and can be considered as Ds 1st with our local attention and NL [15] following after. It can be deduced that $p = 1$ attention receives a down-sampling feature map experiencing information loss, leading to mAP dropping. The next setting for p is 16, but it is inapplicable and not showing in the table because it will change the spatial size of the output feature map, which causes a mismatch with predefined anchors in object detection. Namely, predefined anchors will be four times larger than $p = 16$ because of four times down-sampling in pixel-unshuffle. In sum, local and global attention contribute to each other to form our panel attention. An ideal order in local-to-global makes the contribution positive; $p = 4$ lets local information involve global learning without losing information. To preserve the applicability of panel attention in the current label assignment scheme, we opt to choose $p = 4$ by local-to-global attention order in #7 as the default setting, along with its merit of avoiding information loss, leading to better performance.

B. MEIS

Following the same experimental setting as PASCAL 2012 SBD but with 22k iteration, this simulation is trained end-to-end on MEIS. It represents the real-world clinic scene and tests the backbone capability in detecting big objects, which means owning receptive field ability shall perform well. Recorded results do not include measurement times in FPS in this discussion. Please see Table IV.

TABLE IV:
MEIS results.

#	Work	Backbone	Avg-mAP↑	Mask-mAP↑	Box-mAP↑	FPS(RTX 4090)↑	Size(M)↓
1	YOLACT	ResNet101	-	-	-	-	51.41
2*	maYOLACT	ResNet101	44.74	41.40	48.08	39.91	51.41
3	maYOLACT	CSP-DarkNet	46.18	42.11	50.25	34.65	45.40
			+3.22%	+1.71%	+4.51%	-13.18%	-11.69%
4	maYOLACT	ViT	37.96	38.09	37.83	**63.80**	93.83
		(Base, batch=16)	-15.15%	-14.86%	-15.44%	+42.60%	+82.51%
5	maYOLACT	Swin Transformer	42.68	40.00	45.36	60.00	94.17
		(Base, batch=4)	-4.60%	-3.38%	-5.66%	+50.34%	+83.17%
6	RTMDet	CSP-DarkNet	44.75	40.18	49.33	56.22	80.37
	(ins-X)		+0.02%	-2.95%	+2.60%	+40.87%	+56.33%
7	RAMEM	UPANet80 V2	46.01	42.29	49.73	63.52	**35.82**
		(Vanilla)	+4.07%	+2.15%	+3.43%	+59.16%	-30.32%
8	RAMEM	UPANet80 V2	46.56	42.52	50.59	61.04	38.36
		(GC, Att 1st)	+4.07%	+2.71%	+5.22%	+52.94%	-25.38%
9	RAMEM	UPANet80 V2	46.32	42.43	50.21	61.21	38.36
		(EA, Att 1st)	+3.53%	+2.49%	+4.43%	+53.37%	-23.30%
10	RAMEM	UPANet80 V2	46.25	42.39	50.11	51.76	39.43
		(A2, Att 1st)	+3.38%	+2.39%	+4.22%	+29.69%	-23.30%
11	RAMEM	UPANet80 V2	-	-	-	-	38.63
		(NL, Att 1st)	-	-	-	-	-24.86%
12	RAMEM	UPANet80 V2	46.26	42.23	50.29	61.21	38.36
		(GC, Ds 1st)	+3.40%	+2.00%	+4.60%	+53.47%	-25.38%
13	RAMEM	UPANet80 V2	46.80	42.59	51.00	61.61	38.36
		(EA, Ds 1st)	+4.60%	+2.87%	+6.07%	+54.37%	-25.38%
14	RAMEM	UPANet80 V2	46.41	41.91	49.79	52.93	39.43
		(A2, Ds 1st)	+3.73%	+1.23%	+3.56%	+32.62%	-23.30%
15	RAMEM	UPANet80 V2	46.71	**43.05**	50.35	59.80	38.62
		(NL, Ds 1st)	+4.40%	+3.99%	+4.72%	+49.84%	-24.88%
16	YOLACT	ResNet101	-	-	-	-	93.60
		(Panel attention)	-	-	-	-	+82.07%
17	maYOLACT	ResNet101	45.84	41.94	49.74	60.66	93.60
		(Panel attention)	+2.46%	+1.30%	+3.45%	+51.99%	+82.07%
18	RAMEM	UPANet80 V2	**46.86**	42.71	**51.01**	57.04	40.28
		(Panel attention)	+4.74%	+3.16%	+6.09%	+42.92%	-21.65%

*The baseline

Big receptive field contributes to efficiency and mAPs in M-mode echocardiography – The results reflect the same pattern as Table II, but the benefits of big receptive field reflect in terms of mAP and FPS. As traditional label assignment based on anchor-based IoU in YOLACT fails to operate, the baseline has become maYOLACT ResNet101. From our proposed dataset MEIS, the influence of the receptive field among a backbone can be amplified, as most of the objects occupy a great portion of pixels, which is what pure CNN-based lacks. This statement toward efficiency can be observed from the FPS index in the table, in which bigger receptive field backbones improve in FPS that can not be seen in II. The possible explanation is that the confidence of predicted boxes affects the post-processing of NMS. Less confidence causes more candidates to process, and vice versa. With NL, a big receptive field enables long-range information learning and connects important pixels globally, concreting the prediction and boosting confidence. All in all, this outcome solidifies the need to equip a big receptive field in M-mode echocardiogram detection in our scheme and the merit of panel attention.

C. Attention Map Explanation

To get a picture of what attention is learned and the difference from CNN, a feature map explainable method Score-CAM [50] has been opted. The sampled maps are extracted from four models in the same maYOLACT scheme: ResNet, CSP-DarkNet, Swin Transformer and panel attention in ResNet101 and UPANet80 V2. Also, stages from stage 2 (the upper row) and stage 3 (the bottom row) for viewing the effect of the deep are sampled to examine the growth of the receptive field as well.

Non-local block helps to connect to the right area and prevent receptive field degradation (Fig. 8) – Observing the sampled responding feature maps from MEIS in Fig. 8, it surprisingly shows that the "non-local" information is transforming from a receptive field to the relationship toward

TABLE V:
Measurement results on MEIS.

#	Measurement (Work)	Backbone	Mean (LV View) MAE	Mean (LV View) MSE	Std (LV View) MAE	Std (LV View) MSE	AoR-Diam MAE	AoR-Diam MSE	LA-Dim MAE	LA-Dim MSE	LVIDd MAE	LVIDd MSE	LVPWd MAE	LVPWd MSE	IVSd MAE	IVSd MSE	LVIDs MAE	LVIDs MSE	LVPWs MAE	LVPWs MSE	IVSs MAE	IVSs MSE	FPS (Time sec) RTX 4090
1	Humans (Manual)	-	0.425	0.378	0.221	0.371	0.568	0.441	0.520	0.460	0.783	1.177	0.124	0.022	0.270	0.099	0.651	0.653	0.245	0.080	0.236	0.094	- (>60)
2-1*	MENN (maYOLACT)	ResNet101	0.197	0.107	0.130	0.172	-	-	-	-	0.191	0.006	0.173	0.039	0.123	0.019	0.478	0.489	0.122	0.020	0.097	0.015	20.54 (0.049)
2-2	AMEM (maYOLACT)	ResNet101	0.146 (0.145) -26.40%	**0.036 (0.035)** -67.29%	0.031 (0.034) -73.85%	0.015 (0.016) -90.70%	0.132 -	0.025 -	0.170 -	0.051 -	0.153 -19.90%	0.045 +650.00%	0.160 -7.51%	0.033 -15.38%	0.140 +13.82%	0.025 +31.58%	0.165 -65.48%	0.060 -87.33%	0.177 +45.08%	0.038 +90.00%	**0.073** -24.72%	0.009 -40.00%	27.03 (0.037) +31.60%
3-1*	MENN (maYOLACT)	ViT (Base, batch=16)	0.314	0.306	0.273	0.426	-	-	-	-	0.468	0.571	0.178	0.041	0.137	0.029	0.861	1.149	0.134	0.025	0.106	0.021	22.29 (0.045)
3-2	AMEM (maYOLACT)	ViT (Base, batch=16)	0.150 (0.150) -52.23%	0.040 (0.041) -86.60%	**0.018 (0.020)** -92.67%	0.016 (0.017) -96.01%	0.137 -	0.029 -	0.165 -	0.045 -	0.135 -71.15%	0.032 -94.40%	0.157 -11.80%	0.031 -24.39%	0.154 +12.41%	0.029 0.00%	0.145 -83.16%	0.079 -93.12%	0.186 +38.81%	0.043 +72.00%	0.124 +16.98%	0.034 +61.90%	29.77 (0.034) +33.56%
4-1*	MENN (maYOLACT)	Swin Transformer (Base, batch=14)	0.186	0.066	0.090	0.068	-	-	-	-	0.239	0.094	0.171	0.039	0.130	0.023	0.358	0.205	0.119	0.018	0.096	0.016	22.15 (0.045)
4-2	AMEM (maYOLACT)	Swin Transformer (Base, batch=4)	0.155 (0.157) -15.59%	0.042 (0.043) -34.85%	0.025 (0.026) -71.11%	0.015 (0.016) -76.47%	0.130 -	0.028 -	0.172 -	0.048 -	0.144 -39.75%	0.044 -53.19%	0.148 -13.45%	0.029 -25.64%	0.148 +13.85%	0.027 +17.39%	0.181 -49.44%	0.074 -63.90%	0.199 +67.23%	0.050 +177.78%	0.120 +25.00%	0.033 +106.25%	31.51 (0.032) +42.26%
5-1*	MENN (RTMDet)	CSP-DarkNet	0.161	0.044	0.049	0.029	-	-	-	-	0.183	0.057	0.187	0.044	0.113	0.016	0.237	0.098	0.155	0.033	0.090	**0.014**	20.81 (0.048)
5-3	AMEM (RTMDet)	CSP-DarkNet	0.147 (0.148) -8.07%	0.036 (0.037) -15.91%	0.034 (0.039) -20.41%	0.016 (0.018) -37.93%	0.142 -	0.028 -	**0.142** -	**0.035** -	0.151 -17.49%	-31.58%	0.177 -5.35%	0.040 -9.09%	0.108 -4.42%	0.015 -6.25%	0.162 -31.65%	0.065 -33.67%	0.204 +31.61%	0.048 +45.45%	0.089 -1.11%	**0.014** 0.00%	32.58 (0.031) +56.56%
6-1*	MENN (RAMEM)	UPANet80 V2 (Vanilla)	0.198	0.098	0.121	0.142	-	-	-	-	0.212	0.068	0.193	0.049	0.104	0.016	0.451	0.412	0.144	0.026	0.087	0.015	23.13 (0.043)
6-2	AMEM (RAMEM)	UPANet80 V2 (Vanilla)	0.152 (0.144) -27.27%	0.046 (0.033) -66.33%	0.044 (0.042) -65.29%	0.036 (0.015) -89.44%	0.138 -	0.029 -	0.218 -	0.135 -	**0.106** -50.00%	**0.023** -66.18%	0.181 -6.22%	0.043 -12.24%	0.103 -0.96%	0.015 -6.25%	0.138 -69.40%	**0.030** -92.72%	0.218 +51.39%	0.061 +134.62%	0.118 +35.63%	0.029 +93.33%	34.99 (0.029) +51.28%
7-1*	MENN (RAMEM)	UPANet80 V2 (NL Ds 1st)	0.189	0.082	0.086	0.086	-	-	-	-	0.276	0.138	0.178	0.042	0.112	0.019	0.331	0.250	0.140	0.025	0.096	0.016	23.76 (0.042)
7-2	AMEM (RAMEM)	UPANet80 V2 (NL Ds 1st)	0.148 (0.144) -23.81%	0.039 (0.032) -60.98%	0.036 (0.035) -59.30%	0.026 (0.012) -86.05%	**0.123** -	**0.023** -	0.196 -	0.103 -	0.125 -54.71%	0.029 -78.99%	0.175 -1.69%	0.041 -2.38%	0.112 0.00%	0.018 -5.26%	0.163 -50.76%	0.042 -83.20%	0.194 +38.57%	0.045 +80.00%	0.096 0.00%	**0.014** -12.50%	32.97 (0.030) +38.76%
8-1*	MENN (maYOLACT)	ResNet101 (Panel attention)	0.175	0.072	0.079	0.091	-	-	-	-	0.181	0.054	0.182	0.043	0.121	0.019	0.336	0.274	0.139	0.026	0.092	0.016	25.06 (0.040)
8-2	AMEM (maYOLACT)	ResNet101 (Panel attention)	0.148 (0.149) -14.86%	0.037 (0.037) -48.61%	0.030 (0.034) -56.96%	**0.013 (0.014)** -84.62%	0.133 -	0.028 -	0.161 -	0.043 -	0.145 -19.89%	0.042 -22.22%	0.180 -1.10%	0.043 0.00%	0.127 +4.96%	0.021 +10.53%	0.169 -49.70%	0.058 -78.83%	0.183 +31.65%	0.043 +65.38%	0.088 -4.35%	0.017 +6.25%	**36.32 (0.028)** +44.93%
9-1*	MENN (RAMEM)	UPANet80 V2 (Panel attention)	0.199	0.163	0.182	0.302	-	-	-	-	0.219	0.080	**0.097**	**0.015**	0.094	**0.014**	0.592	0.836	0.083	**0.010**	0.109	0.025	23.79 (0.042)
9-2	AMEM (RAMEM)	UPANet80 V2 (Panel attention)	**0.137 (0.121)** -39.20%	0.042 (0.026) -84.05%	0.037 (0.016) -91.21%	0.043 (0.011) -96.36%	0.153 -	0.031 -	0.222 -	0.152 -	0.133 -39.27%	0.040 -50.00%	0.114 +17.53%	0.033 +120.00%	**0.015** -84.04%	0.149 +964.29%	**0.042** -92.91%	0.103 -87.68%	**0.014** -83.13%	0.121 +1110.00%	0.121 +11.01%	0.021 -16.00%	31.85 (0.031) +33.88%

*The baseline

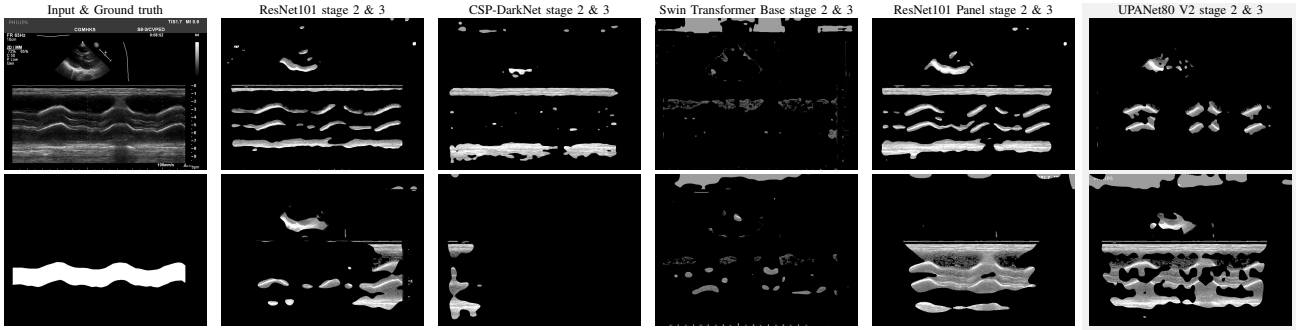

Fig. 8: MEIS sampled feature maps with different backbones. Sampled images only show the responding area where the value is over 0.5 in a [0, 1] range.

ground truth under the method of Score-CAM. In another scene in stage 3, deeper responding layers indicate that the focus will narrow to the related area toward the ground truth. However, what CNNs (ResNet and CSP-DarkNet) fail is the responding area degrading or, even worse, vanishing into a blank and irrelevant area. Evidence examples in Swin Transformer and UPANet show a wider response area, indicating the contribution of global attention. The attention maps reflect the performance behaviour in Table III and Table IV, where the backbones with a big receptive field perform better. With equipping panel attention into the limited receptive field backbone of ResNet, it can be observed that the degradation has been addressed. Moreover, the receptive field focuses on the right place toward the ground truth that happens in ResNet101 Panel and UPANet80 V2. The observation explains the results of NLs and Transformers in Table III and Table IV.

D. AMEM in MEIS

Taking the trained scheme from Table IV with the proposed automatic measurement algorithm AMEM, RAMEM shapes a scheme that can contribute to daily clinic cardiac examination in M-mode echocardiogram. To evaluate the legitimacy of the proposed scheme, the 27 patients' examinations are collected from clinical visits in Table V, which are contained in the testing data and have a total of 40 human testing data because some may have a single view. To make a fair comparison, we gathered 20 human testing data for each point. Also, the sampled images in AV are in Fig. 9. LV follows the same in Fig. 10. Among the sampled images, ground truth images show a possible location of diastole and systole. The key is to locate one of the period locations to get the indicators. The sampled results have been implemented along with the original results on the top row and zoom-in results on the bottom. The compared candidates include the practical manual examination results from Humans, another existing automatic measurement toward LV from MENN, and our proposed method AMEM. The error evaluation index in Table V covers mean absolute error (MAE), mean square error (MSE), and costing time (FPS). The mean and standard deviation toward each error index and indicator are presented to make a comprehensive comparison even more effortless.

AMEM upon UPANet80 V2 (RAMEM) surpasses hu-

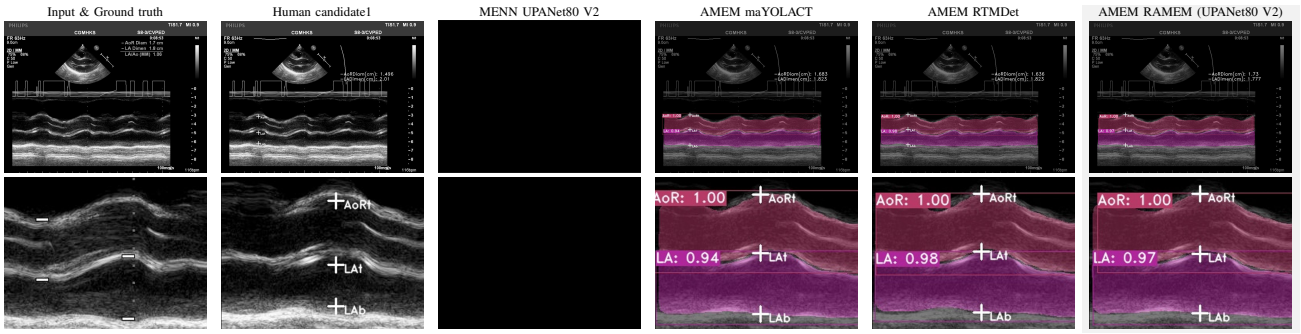

Fig. 9: AV sampled measurement results. The t indicates the top of the boundary, and the b indicates the bottom.

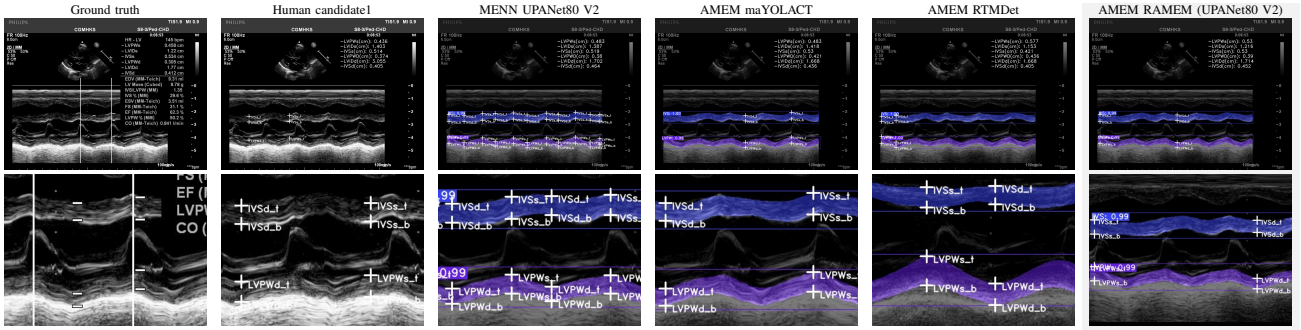

Fig. 10: LV sampled measurement results. As there are diastole and systole moments, each moment is abbreviated into s and d, along with the same top and bottom boundary abbreviations as the above figures. The locations may vary since AMEM takes the topmost point as the systole and the lowest point as the diastole.

man and MENN in real-time detection (Table V. Fig. 9 and Fig. 10) – The results show two milestones: 1) seamless diagnosis by making the whole scheme in real-time under AMEM, and 2) minimizing mean and variance of error while surpassing human performance, MENN, and other backbones. Apart from that, AMEM is able to measure AV view, which MENN fails to do. The sampled figures further solidfy our statement of AMEM superiority by showing a more accurate anchor on the mask. As MENN inclines to calculate the average length from each diastole and systole point, the cost time and detection are jeopardized by the number of anchors and in-accurate points. Although such measurement covers all possible points that can prevent outlier effects, a poor sample period could cause the result to be sensitive to a specific scenario. When comparing with the results from human candidate 1, biased locating has occurred. The bias happened at systole and diastole or vessel boundary. Conversely, the sampled measurement results from AMEM have accurately located end-systole in AV, even in less apparent echocardiograms. The same result is also shown in LV with less bias. In Table V, we state our proposed scheme in #9-2 still possesses the best candidate because of two-fold reasons: 1) comparing the possessing number of the best indices, #9 occupies the most, which is aligned with the best average mAP in Table IV; 2) picking among the group #9, #9-2 has better performance in terms of fewer mean & standard deviation error and real-time capability. Therefore, we opt #9-2 as our RAMEM scheme.

VI. Conclusion

This study introduces solutions for challenges in M-mode echocardiography with real-time detection. The MEIS dataset links RIS and M-mode echocardiography, enabling the development of an automatic detection system. Panel attention addresses the identification of large objects in M-mode echocardiograms, while AMEM demonstrates impressive performance with minimal errors and real-time capabilities, making it a strong candidate for routine clinical use. These innovations aim to advance medical imaging and computer vision, particularly in echocardiography.

References

[1] C. M. Otto, *Textbook of clinical echocardiography*. Elsevier Health Sciences, 2013.

[2] E. Cavero, A. Alesanco, and J. García, "Real-time echocardiogram transmission protocol based on regions and visualization modes," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 5, pp. 1668–1677, 2014.

[3] E. Cavero, A. Alesanco, L. Castro, J. Montoya, I. Lacambra, and J. Garcia, "Spiht-based echocardiogram compression: clinical evaluation and recommendations of use," *IEEE Journal of Biomedical and Health Informatics*, vol. 17, no. 1, pp. 103–112, 2012.

[4] M. S. Sultan, N. Martins, E. Costa, D. Veiga, M. J. Ferreira, S. Mattos, and M. T. Coimbra, "Virtual m-mode for echocardiography: A new approach for the segmentation of the anterior mitral leaflet," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 1, pp. 305–313, 2018.

[5] X. Cui, Y. Cao, Z. Liu, X. Sui, J. Mi, Y. Zhang, L. Cui, and S. Li, "Trsa-net: Task relation spatial co-attention for joint segmentation, quantification and uncertainty estimation on paired 2d echocardiography," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 8, pp. 4067–4078, 2022.

[6] W. Xue, H. Cao, J. Ma, T. Bai, T. Wang, and D. Ni, "Improved segmentation of echocardiography with orientation-congruency of optical flow and motion-enhanced segmentation," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 12, pp. 6105–6115, 2022.

[7] N. Painchaud, N. Duchateau, O. Bernard, and P.-M. Jodoin, "Echocardiography segmentation with enforced temporal consistency," *IEEE Transactions on Medical Imaging*, vol. 41, no. 10, pp. 2867–2878, 2022.

[8] K. B. Girum, G. Créhange, and A. Lalande, "Learning with context feedback loop for robust medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 40, no. 6, pp. 1542–1554, 2021.

[9] S. Leclerc, E. Smistad, J. Pedrosa, A. Østvik, F. Cervenansky, F. Espinosa, T. Espeland, E. A. R. Berg, P.-M. Jodoin, T. Grenier, *et al.*, "Deep learning for segmentation using an open large-scale dataset in 2d echocardiography," *IEEE Transactions on Medical Imaging*, vol. 38, no. 9, pp. 2198–2210, 2019.

[10] C. Duan, M. K. Montgomery, X. Chen, S. Ullas, J. Stansfield, K. McElhanon, and D. Hirenallur-Shanthappa, "Fully automated mouse echocardiography analysis using deep convolutional neural networks," *American Journal of Physiology-Heart and Circulatory Physiology*, vol. 323, no. 4, pp. H628–H639, 2022.

[11] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "Yolact: Real-time instance segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9157–9166, 2019.

[12] D. Bolya, C. Zhou, F. Xiao, and Y. Lee, "Yolact++: Better real-time instance segmentation. arxiv 2019," *arXiv preprint arXiv:1912.06218*.

[13] Y. Lee and J. Park, "Centermask: Real-time anchor-free instance segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13906–13915, 2020.

[14] C. Lyu, W. Zhang, H. Huang, Y. Zhou, Y. Wang, Y. Liu, S. Zhang, and K. Chen, "Rtmdet: An empirical study of designing real-time object detectors," *arXiv preprint arXiv:2212.07784*, 2022.

[15] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7794–7803, 2018.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[17] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*, pp. 6105–6114, PMLR, 2019.

[18] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[19] N. Park and S. Kim, "How do vision transformers work?," *arXiv preprint arXiv:2202.06709*, 2022.

[20] S. Lee, S. Lee, and B. C. Song, "Improving vision transformers to learn small-size dataset from scratch," *IEEE Access*, vol. 10, pp. 123212–123224, 2022.

[21] C.-H. Tseng, S.-J. Lee, J. Feng, S. Mao, Y.-P. Wu, J.-Y. Shang, and X.-J. Zeng, "Upanets: Learning from the universal pixel attention neworks," *Entropy*, vol. 24, no. 9, p. 1243, 2022.

[22] G. Zamzmi, L.-Y. Hsu, W. Li, V. Sachdev, and S. Antani, "Harnessing machine intelligence in automatic echocardiogram analysis: Current status, limitations, and future directions," *IEEE Reviews in Biomedical Engineering*, vol. 14, pp. 181–203, 2020.

[23] A. Ghorbani, D. Ouyang, A. Abid, B. He, J. H. Chen, R. A. Harrington, D. H. Liang, E. A. Ashley, and J. Y. Zou, "Deep learning interpretation of echocardiograms," *NPJ Digital Medicine*, vol. 3, no. 1, p. 10, 2020.

[24] D. Ouyang, B. He, A. Ghorbani, N. Yuan, J. Ebinger, C. P. Langlotz, P. A. Heidenreich, R. A. Harrington, D. H. Liang, E. A. Ashley, *et al.*, "Video-based ai for beat-to-beat assessment of cardiac function," *Nature*, vol. 580, no. 7802, pp. 252–256, 2020.

[25] G. Zamzmi, S. Rajaraman, L.-Y. Hsu, V. Sachdev, and S. Antani, "Real-time echocardiography image analysis and quantification of cardiac indices," *Medical Image Analysis*, vol. 80, p. 102438, 2022.

[26] A. Madani, R. Arnaout, M. Mofrad, and R. Arnaout, "Fast and accurate view classification of echocardiograms using deep learning," *NPJ digital medicine*, vol. 1, no. 1, p. 6, 2018.

[27] C. Fancourt, K. Azer, S. L. Ramcharan, M. Bunzel, B. R. Cambell, J. R. Sachs, and M. Walker, "Segmentation of arterial vessel wall motion to sub-pixel resolution using m-mode ultrasound," in *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 3138–3141, IEEE, 2008.

[28] K. Oksuz, B. C. Cam, F. Kahraman, Z. S. Baltaci, S. Kalkan, and E. Akbas, "Mask-aware iou for anchor assignment in real-time instance segmentation," *arXiv preprint arXiv:2110.09734*, 2021.

[29] S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li, "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9759–9768, 2020.

[30] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "Yolox: Exceeding yolo series in 2021," *arXiv preprint arXiv:2107.08430*, 2021.

[31] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11976–11986, 2022.

[32] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10012–10022, 2021.

[33] H. Hu, Z. Zhang, Z. Xie, and S. Lin, "Local relation networks for image recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3464–3473, 2019.

[34] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens, "Stand-alone self-attention in vision models," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[35] M.-H. Guo, C.-Z. Lu, Z.-N. Liu, M.-M. Cheng, and S.-M. Hu, "Visual attention network," *arXiv preprint arXiv:2202.09741*, 2022.

[36] L. Zhou, H. Cai, J. Gu, Z. Li, Y. Liu, X. Chen, Y. Qiao, and C. Dong, "Efficient image super-resolution using vast-receptive-field attention," in *European Conference on Computer Vision*, pp. 256–272, Springer, 2022.

[37] X. Li, Z. Zhong, J. Wu, Y. Yang, Z. Lin, and H. Liu, "Expectation-maximization attention networks for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9167–9176, 2019.

[38] Z. Geng, M.-H. Guo, H. Chen, X. Li, K. Wei, and Z. Lin, "Is attention better than matrix decomposition?," *arXiv preprint arXiv:2109.04553*, 2021.

[39] Y. Chen, Y. Kalantidis, J. Li, S. Yan, and J. Feng, "Aˆ 2-nets: Double attention networks," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[40] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "Gcnet: Non-local networks meet squeeze-excitation networks and beyond," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) workshops*, pp. 0–0, 2019.

[41] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7132–7141, 2018.

[42] M.-H. Guo, Z.-N. Liu, T.-J. Mu, and S.-M. Hu, "Beyond self-attention: External attention using two linear layers for visual tasks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[43] B. Wandt, L. Bojö, K. Tolagen, and B. Wranne, "Echocardiographic assessment of ejection fraction in left ventricular hypertrophy," *Heart*, vol. 82, no. 2, pp. 192–198, 1999.

[44] K. Mizukoshi, M. Takeuchi, Y. Nagata, K. Addetia, R. M. Lang, Y. J. Akashi, and Y. Otsuji, "Normal values of left ventricular mass index assessed by transthoracic three-dimensional echocardiography," *Journal of the American Society of Echocardiography*, vol. 29, no. 1, pp. 51–61, 2016.

[45] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision – ECCV 2014*, pp. 740–755, Springer, 2014.

[46] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1251–1258, 2017.

[47] S. Suzuki *et al.*, "Topological structural analysis of digitized binary images by border following," *Computer Vision, Graphics, and Image Processing*, vol. 30, no. 1, pp. 32–46, 1985.

[48] J. Sklansky, "Finding the convex hull of a simple polygon," *Pattern Recognition Letters*, vol. 1, no. 2, pp. 79–83, 1982.

[49] H. Zhu, B. Chen, and C. Yang, "Understanding why vit trains badly on small datasets: An intuitive perspective," *arXiv preprint arXiv:2302.03751*, 2023.

[50] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, and X. Hu, "Score-cam: Score-weighted visual explanations for convolutional neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) workshops*, pp. 24–25, 2020.