# Improving the Quality of Fetal Heart Ultrasound Imaging With Multihead Enhanced Self-Attention and Contrastive Learning

Yingying Zhang, *Student Member, IEEE*, Haogang Zhu, Jian Cheng, Jingyi Wang, Xiaoyan Gu, Jiancheng Han, Ye Zhang, Ying Zhao, Yihua He, and Hongjia Zhang

***Abstract*—Fetal congenital heart disease (FCHD) is a common, serious birth defect affecting ∼1% of newborns annually. Fetal echocardiography is the most effective and important technique for prenatal FCHD diagnosis. The prerequisites for accurate ultrasound FCHD diagnosis are accurate view recognition and high-quality diagnostic view extraction. However, these manual clinical procedures have drawbacks such as, varying technical capabilities and inefficiency. Therefore, the automatic identification of high-quality multiview fetal heart scan images is highly desirable to improve prenatal diagnosis efficiency and accuracy of FCHD. Here, we present a framework for multiview fetal heart ultrasound image recognition and quality assessment that comprises two parts: a multiview classification and localization network (MCLN) and an improved contrastive learning network (ICLN). In the MCLN, a multihead enhanced self-attention mechanism is applied to construct the classification network and identify six accurate and interpretable views of the fetal heart. In the ICLN, anatomical structure standardization and image clarity are considered. With contrastive learning, the absolute loss, feature relative loss and predicted value relative loss are combined to achieve favorable quality assessment results. Experiments show that the MCLN outperforms other state-of-the-art networks by 1.52–13.61% when determining the F1 score in six standard view recognition tasks, and the ICLN is comparable to the performance of expert cardiologists in the quality assessment of fetal heart ultrasound images, reaching 97% on a test set within 2 points for the four-chamber view task. Thus, our architecture offers great potential in helping cardiologists improve quality control for fetal echocardiographic images in clinical practice.**

***Index Terms*—Contrastive learning, fetal congenital heart disease, fetal echocardiography, view recognition, quality assessment, multihead enhanced self-attention.**

Manuscript received 2 March 2023; revised 23 June 2023; accepted 2 August 2023. Date of publication 9 August 2023; date of current version 7 November 2023. This work was supported in part by the National Key R&D Program of China under Grant 2021ZD0140407, in part by the National Natural Science Foundation of China under Grant U21A20523, in part by the Beijing Natural Science Foundation under Grant L222152, in part by the Beijing Municipal Administration of Hospitals Incubating Program under Grant PX2022026, in part by the Fundamental Research Funds for the Central Universities, Beijing Advanced Innovation Center for Big Data-Based Precision Medicine, Beihang University, Beijing, China, in part by the State Key Laboratory of Software Development Environment, Beihang University, Beijing, China, Beijing Lab for Cardiovascular Precision Medicine under Grant PXM2018_014226_000013, and in part by the Beijing Key Laboratory of Maternal-Fetal Medicine in Fetal Heart Disease under Grant BZ0308. *(Corresponding authors: Haogang Zhu; Yihua He; Hongjia Zhang.)*

Yingying Zhang is with the School of Biological Science and Medical Engineering, Beihang University, Beijing 100191, China (e-mail: zhangyingying@buaa.edu.cn).

Haogang Zhu and Jian Cheng are with the State Key Laboratory of Software Development Environment, School of Computer Science and Engineering, Beihang University, Beijing 100191, China, and also with the Zhongguancun Laboratory, Beijing 102206, China (e-mail: haogangzhu@buaa.edu.cn; jian_cheng@buaa.edu.cn).

Jingyi Wang, Xiaoyan Gu, Jiancheng Han, Ye Zhang, Ying Zhao, and Yihua He are with the Echocardiography Medical Center, Beijing Anzhen Hospital, Capital Medical University, Beijing 100054, China (e-mail: wangjingyianzhen@163.com; xiaoyan_gu@yahoo.com; han_jc1977@hotmail.com; zhang3389@qq.com; yingzhaoecho@163.com; heyihuaecho@hotmail.com).

Hongjia Zhang is with the Beijing Lab for Cardiovascular Precision Medicine, Beijing 100029, China (e-mail: zhanghongjia722@hotmail.com).

Digital Object Identifier 10.1109/JBHI.2023.3303573

## I. Introduction

Fetal congenital heart disease (FCHD) is a common, and serious congenital malformation worldwide and is the greatest birth defect-related contributor to infant mortality [1], [2], [3]. Recently, the global prevalence of FCHD at birth has been 9.4‰, and the reported prevalence of FCHD is increasing [3], [4]. Early diagnosis of FCHD is essential to improving the prognosis [5]. Fetal echocardiography is the most effective and important technique for the prenatal diagnosis of FCHD [6]. The prerequisite for accurate ultrasound diagnosis of the fetal heart is accurate view recognition and the identification of high-quality diagnostic views [7], [8], [9], [10]. Clinically, the view recognition and quality assessment used to diagnose FCHD is usually performed manually, which has some drawbacks. For example, the process is labor intensive and subjective, technical capabilities vary greatly and it has low efficiency. As shown in Fig. 1, fetal echocardiography has six guideline-recommended standard views for the diagnosis of FCHD, including the four-chamber view (4CV), the left and right ventricular outflow tract (LVOT, RVOT) views, the three-vessel view (3VV), the three-vessel and trachea (3VT) view, and the transverse abdominal view (TAV) [11], [12]. The combination of multiple views can greatly improve the detection rate and diagnostic accuracy of FCHD [13], [14], [15]. However, based on the commonly used 4CV scan, each extra view takes at least twice the time corresponding to each 4CV scan [13]. Moreover, the acquisition of multiple standard views requires sonographers to master the spatial correspondence betweenthe three-dimensional structure

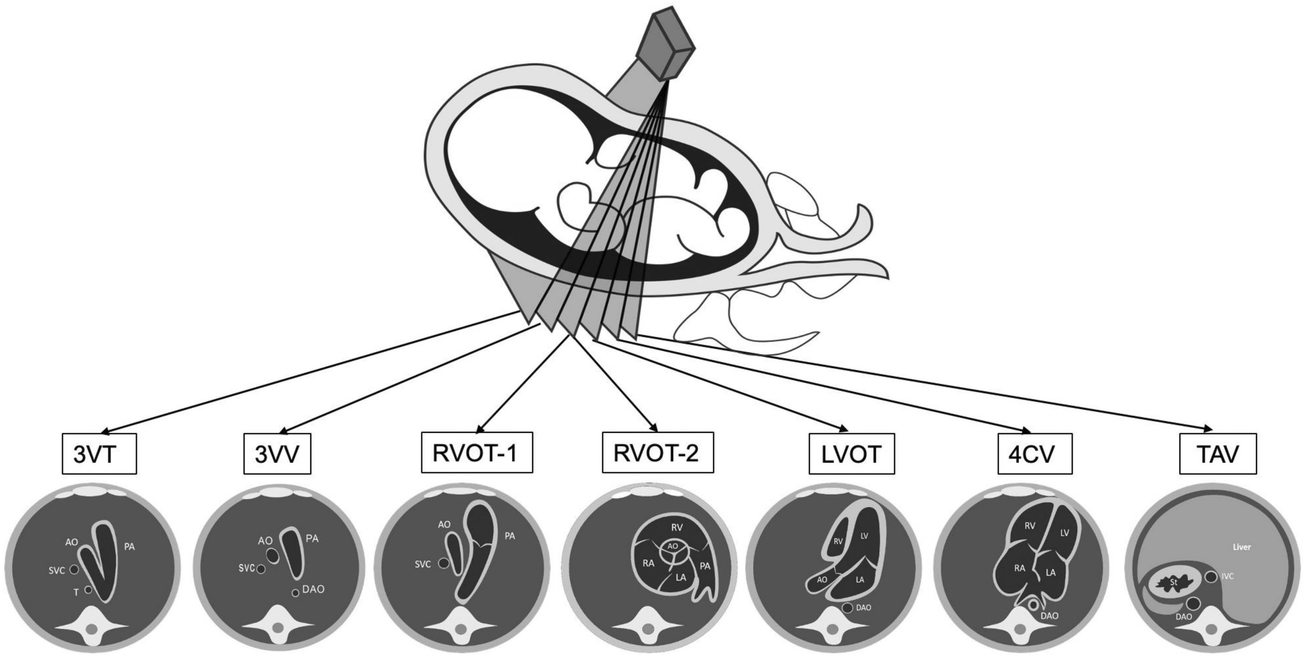

Fig. 1. Six fetal cardiac ultrasound image views, where the main anatomical structures comprise the stomach bubble (St), descending aorta (DAO), liver and inferior vena cava (IVC) in TAV; left atrium (LA), left ventricle (LV), right atrium (RA), right ventricle (RV) and descending aorta (DAO) in 4CV; left atrium (LA), left ventricle (LV), right ventricle (RV), aortic root (Ao) and descending aorta (DAO) in LVOT; pulmonary artery (PA), superior vena cava (SVC) and aorta (AO) in RVOT-1; right ventricle (RV), pulmonary artery (PA), left atrium (LA), right atrium (RA) and aorta (AO) in RVOT-2; pulmonary artery (PA), aorta (AO), superior vena cava (SVC) and descending aorta (DAO) in 3VV; and pulmonary artery (PA), aorta (AO), superior vena cava (SVC) and trachea (Tr) in 3VT.

of the heart and the two-dimensional ultrasound views while obtaining the standard views required for diagnosis within a limited examination time and acoustic window. This capacity is difficult to acquire with short training sessions, and centers that can perform the training are scarce. Overall, there are very few physicians who can master fetal cardiac examination of multiple views for FCHD; such expertise cannot be cultivated in a short period of time but is urgently needed. Therefore, exploring artificial intelligence (AI) inference models and methods that can address automatic view recognition and quality assessment in fetal heart ultrasound scans is highly desirable.

With the development of AI techniques, an abundance of studies on echocardiographic images have recently been conducted to optimize the scanning process, obtain the standard view and enhance image quality [16], [17], [18], [19]. For instance, Narang et al. [20] used AI techniques to guide nurses without any ultrasound experience to successfully obtain 10 echocardiographic views of diagnostic value. Wu et al. [21] proposed a deep learning network based on knowledge distillation to identify 23 standard echocardiographic views commonly used automatically and effectively for the diagnosis of congenital heart disease in children and obtained a good recognition effect. Abdi et al. [22] classified the image quality of adult four-chamber views into five grades, assessed the image quality by adding noise for distortion simulation, and then used a deep neural network to perform regression prediction of echocardiographic image quality. Thus, AI is an effective tool for echocardiographic image analysis. In contrast to the adult heart, fetal heart ultrasound images need to be collected through the mother's uterus. The fetal heart has the characteristics of being small in size and fast beating and has different and changing positions with different gestational ages, which increases the difficulty of fetal heart image analysis.

Recently, there have been several attempts at quality control of fetal echocardiography, which can be roughly divided into structure-based, and image-clarity based methods [23], [24], [25]. However, the current state of the research has several limitations. First, a structure-based quality assessment network gives scores based on the results identified by the object detection network. The target of network optimization is to identify as much of this view as possible, even if a substructure is missing or unclear. This is contrary to the optimization of clinical applications. Clinical fetal heart examination is a video sweep, and the same view has a large number of video frames. We hope to extract the standard plane with the most complete structure and the clearest image quality for diagnosis. In addition, a structure-based quality assessment network requires experts to annotate the amount of structure in each view to meet the high data volume requirements of the detection network, which is a repetitive and laborious manual operation with a heavy labeling workload. Second, image quality evaluation methods based on image clarity ignore the importance of the anatomical structure, and distortion simulations of ultrasound image quality, which consider factors such as blur, image compression and other artificially added distortions, are not applicable to medical images. Finally, most of the studies are based on a single view and a single task, and there is no whole-process research to address the problem of view extraction and quality assessment from the clinical reality of the cardiac panorama. Therefore, the difficulties we face can be divided into the following two points.

First, we must synchronously achieve view identification and quality scoring in one process. Second, images with the same score are derived from different features, including different maternal bodies, different gestational ages, and different positions. In this case, for the desired scope, a structured method to process and project all the different features into a unique hidden representation is needed.

In this article, we consider both the anatomical structure of the fetal heart and the image clarity to address the view classification and quality assessment of fetal heart views in a complete method, realize the automatic view recognition and quality assessment of the six views of the fetal heart and improve the diagnostic efficiency and accuracy of FCHD in clinical practice. Furthermore, automatic view recognition and quality assessment of fetal echocardiographic images also provide the basis for automatic diagnosis of FCHD.

In summary, the contributions of this article are as follows:

First, we introduce multihead enhanced self-attention (MESA) mechanisms into the analysis framework for fetal cardiac ultrasound scans to increase the structural focus in the view identification process and improve the view classification accuracy.

Second, we propose a multiple iterative regression strategy to locate the optimal bounding box, achieving an outcome equivalent to that of strongly supervised learning models. This approach allows us to identify crucial anatomical structure regions in fetal heart ultrasound images, eliminate interference regions, and acquire images containing only the areas-of-interest for performing quality assessment.

Third, we propose an improved contrastive learning network with a new strategy and optimize the characteristics of the absolute space that are not well measured by the relative space metric to provide a more reasonable score, and better evaluate the quality of the fetal heart ultrasound images. In addition, the loss function incorporates the relative loss of the input pair of samples, thus further constraining the distribution of the sample features.

## II. Related Work

### A. Automated Fetal Echocardiography Analysis

Fetal echocardiography enables the collection of detailed information on a baby's heart before birth; however, the fetal heart is small and beats fast in utero, and its structure changes with gestational age, which makes it difficult to analyze fetal echocardiography automatically. Moreover, in contrast to other imaging modalities, ultrasound devices cannot automatically acquire images. The data acquisition process heavily depends on the sonographer's knowledge of the fetal heart. Therefore, the quality of fetal echocardiographic images varies considerably, which has a substantial impact on FCHD diagnoses.

Recently, convolutional neural networks (CNNs) have shown good performance in automatic analyses of fetal echocardiography data, including optimizing the scanning process, obtaining standard views, and assessing image quality. For example, Yang et al. [26] used postmortem fetal heart and cardiovascular casts combined with CT scans and fetal echocardiography data to determine the pose relationship between each section and optimize the data acquisition process. Chen et al. [25] explored a neural network to automatically detect standard views from fetal heart ultrasound scan videos and selected different views by using a threshold value. Baumgartner et al. [27] proposed a CNN method called SonoNet to detect 13 standard views of fetuses in 2D ultrasound data and locate the key anatomical structures on the plane. Structure-specific quality assessment has also been investigated through the detection of fetal echocardiographic structures. Dong et al. [23] proposed an automated quality control framework based on structural detection for achieving the identification of a standard 4CV scan for fetal cardiac ultrasounds. Wu et al. [24] developed a deep convolutional neural network to evaluate the image quality of the fetal abdominal plane and regressed the image quality score by detecting key anatomical features of the stomach bubble and umbilical vein. However, previous studies focused on only single views of the fetal heart or single tasks, which differs from actual clinical decision-making processes. In addition, the accuracy of identification of multiple views of the fetal heart is too low to be suitable for clinical application.

Quality assessment networks for fetal echocardiography images usually model quality scoring as a classification or regression problem and input data to learn the mapping between the input and the output [28], [29]. This method is simple and intuitive but ignores the psychological cognitive mechanism of humans. In cognitive psychology, it is asserted that there is actually a relative aesthetic mechanism in the process of human aesthetics [30], and it is similar in the quality scoring of images, which is called the relative scoring mechanism. This means that when humans score images, they tend to refer to other images and give a score by comparing the images with each other. The scoring process is more like a ranking process rather than a process that directly produces an absolute score. Each image $x_i$, has a reference image $x_j$. If the quality of $x_i$ is considerably better than that of $x_j$, $x_i$ is given a higher score; if the quality of $x_i$ is notably worse than that of $x_j$, $x_i$ is given a lower score. Therefore, humans score by comparing two images with each other. If we simply input an image into a CNN and model it as a classification or regression problem, we ignore this relative scoring mechanism. For the modeling of the relative scoring mechanism, the input is usually sample pairs, and then the features are extracted for comparison.

### B. Attention Mechanism

Attention mechanisms have been involved in the human visual system as a way to direct attention to the most important regions while ignoring irrelevant parts. It has been successfully used in natural language processing tasks. This has also inspired researchers to introduce attention mechanisms into computer vision systems to improve the performance of image processing. Self-attention, as an effective attention mechanism, was first proposed in [31] and rapidly provided great advances in various fields, which correlates different locations to calculate an interactive representation for long-range dependency modeling. Wang et al. [32] first introduced self-attention to computer vision tasks and demonstrated great success in video understanding and object detection with nonlocal modeling. Zhang et al. [33] proposed SAGAN, which incorporates a self-attention mechanism into the generative adversarial framework, achieving better

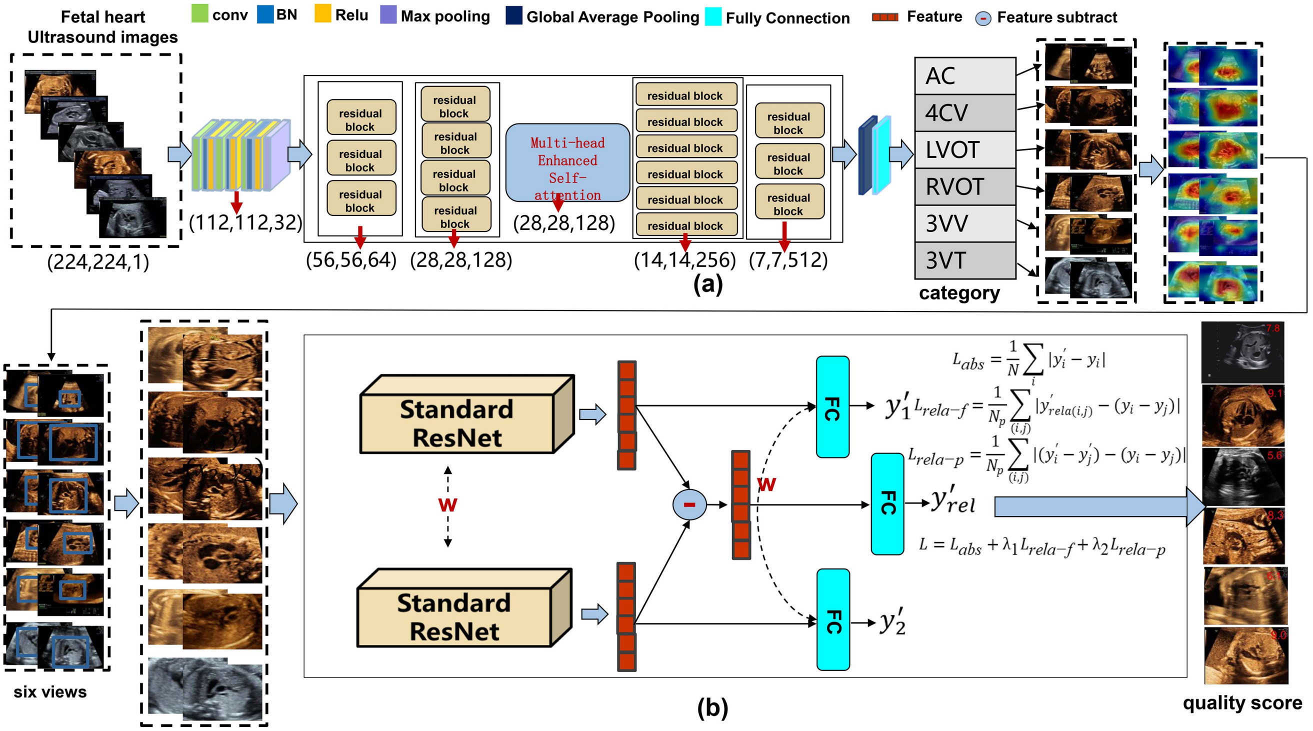
Fetal heart Ultrasound images
conv
BN
Relu
Max pooling
Global Average Pooling
Fully Connection
Feature
Feature subtract
(224,224,1)
(112,112,32)
residual block
(56,56,64)
(28,28,128)
Multi-head Enhanced Self-attention
(28,28,128)
(14,14,256)
(7,7,512)
(a)
AC
4CV
LVOT
RVOT
3VV
3VT
category
six views
Standard ResNet
W
Standard ResNet
W
FC
$y'_1$
FC
$y'_{rel}$
FC
$y'_2$
$L_{abs} = \frac{1}{N}\sum_i |y'_i - y_i|$
$L_{rela-f} = \frac{1}{N_p}\sum_{(i,j)} |y'_{rela(i,j)} - (y_i - y_j)|$
$L_{rela-p} = \frac{1}{N_p}\sum_{(i,j)} |(y'_i - y'_j) - (y_i - y_j)|$
$L = L_{abs} + \lambda_1 L_{rela-f} + \lambda_2 L_{rela-p}$
7.8
9.1
5.6
8.3
6.1
9.0
(b)
quality score

Fig. 2. Proposed architecture for fetal heart multiview recognition and quality assessment. The proposed approach consists of a multiview classification and localization network (MCLN) and an improved comparative learning network (ICLN). MCLN is presented in (a), while ICLN is presented in (b).

generation details with higher parameter efficiency. Unlike self-attention, multihead attention uses the information from multiple subspaces at different positions and acquires the short- and long-range dependence of each single-head attention jointly to enhance the important image features. Zhang et al. [34] proposed an improved version of multihead self-attention called multihead enhanced self-attention (MESA) and achieved impressive feature extraction and reconstruction results in one-class classification experiments. However, whether MESA can be applied in a multiview recognition network based on fetal heart ultrasound images remains to be determined. In this work, we employed a multihead enhanced self-attention mechanism incorporated in a deep residual classification network to improve the classification accuracy and interpretability.

### C. Contrastive Learning

Contrastive learning focuses on learning the common features between similar instances and distinguishing the differences between nonsimilar instances. The input to the contrast network is usually from different images, and the comparability is determined by the pretext task. The pretext task usually involves some rules that define which images are similar or not, thus providing supervision to train the model. This methodology has been widely used in vision tasks and has achieved promising performance. For instance, Wu et al. [35] introduced contrastive learning to an object detection network to detect smoke images, which aims to obtain the internal consistency between augmented images of the same smoke image. Wang et al. [36] employed contrastive learning in a histopathological image classification network to learn global and local image representations. Sun et al. [37] used contrastive learning for a thyroid nodule classification network to improve the accuracy of diagnosis and specificity of biopsy recommendations. However, these contrast learning networks perform random augmentation, such as flipping, rotating, and cropping, on the original image to generate input pairs, which may lead to too low or too high input image similarity and are difficult to directly apply to the analysis of fetal heart ultrasound images. In this article, we propose an improved contrastive learning network to evaluate the quality of fetal heart ultrasound images using a new contrast strategy and optimizing the features of the absolute space by the relative spatial metric. To the best of our knowledge, we are the first to explore contrastive learning for the quality assessment of fetal heart ultrasound images.

## III. Methodology

The pipeline of the proposed framework is illustrated in Fig. 2. First, we propose a multiview classification and localization network (MCLN) to realize view recognition and key region extraction. The MESA mechanism enables the MCLN to focus more on target information, suppress other information, and locate areas of interest in the image for quality assessment. Second, we propose an improved contrastive learning network (ICLN) to evaluate the quality of recognized and extracted images in each view. Specifically, the ICLN uses the input image pair to learn the absolute and relative scores. The network considers not only the absolute loss but also the relative loss of the input sample pairs. When the absolute loss is difficult to optimize, the distribution of the sample features can be further constrained by the relative loss. According to the output of the two networks, the overall quantitative score and the view category of each image are obtained.

## A. Multiview Recognition and Localization

The transformation of multiple views during fetal heart ultrasound scans is very rapid, and the manual extraction of views and locations of anatomical structures is very time-consuming and labor intensive. Therefore, automatic view recognition and extraction is a key step for FCHD diagnosis. In addition to the cardiac structure, ultrasound images usually contain other information, such as artifacts, speckle noise, noncardiac structures, and various parameters marked by sonographers, which are not valuable for quality assessment. Due to different parameters, such as zoom settings, there are also large individual differences in the proportion of the fetal heart structure in the overall image. However, FCHD diagnosis mainly focuses on the cardiac region and does not pay much attention to other areas. Therefore, extracting the anatomical structural regions of the fetal heart ultrasound images that we require, instead of the noise components, is crucial.

The MESA mechanism can capture both short-range and long-range correlations in various subspaces, allowing the model to concentrate on specific features that carry essential information about the input image. Moreover, it can merge the correlations to enhance important correlations while suppressing unimportant correlations. Therefore, this attention mechanism enables the view classification results to focus on the most important structural differences instead of the interference area. Moreover, due to its ability to focus on the key regions, from coarse to fine levels of detail, we extract only the key areas to score the quality and ignore the interference regions. Specifically, the class activation map reflects the activation of a certain class in the feature map of a CNN. It is typically used for the visualization of the neural network feature map and can also be used for weak supervision target location [38]. The quality assessment of fetal heart ultrasound images mainly focuses on the anatomical structure in each view while paying no attention to other background information during scoring. Too much background information may affect the scoring results, so it is necessary to identify the fetal heart region for a more accurate quality evaluation.

To generate surrounding boxes for the heart region, we employed MESA-enhanced class activation maps. However, the accuracy of the generated boxes was challenging to determine due to the difficulty in setting appropriate threshold values for the class activation maps. The surrounding boxes were considered accurate if they covered the entire heart structure region with minimal background information. Conversely, surrounding boxes were considered inaccurate if they did not fully cover the heart region or were excessively large. The pseudo supervised object localization (PSOL) network proposed by Zhang et al. improves the object localization accuracy by using a regression model to regress the generated pseudo bounding boxes [39]. Inspired by the PSOL network, we propose a multiple iterative regression strategy to select the surrounding boxes for the fetal heart region. This strategy involves conducting several regression optimizations before training the final localization model using pseudo bounding boxes. The dataset is randomly divided into two (or more) mutually exclusive sets, and one set is used to train the model to predict the other set. If the pseudo bounding box predicted for the other set deviates considerably from the original box (i.e., a small IoU between the predicted and original boxes), the predicted pseudo bounding box is used as the new pseudo bounding box. The specific algorithm is shown in Algorithm 1.

**Algorithm 1:** Multiple Iterative Regression Strategy.

```
INPUT: Training set T, threshold $\gamma$
1. The training set $T$ is randomly divided into $A$ and B,
   where $A$ and $B$ represent different cases;
2. Model $M_A$ is trained based on training set $A$, and
   model $M_B$ is trained based on training set $B$;
3. Model $M_A$ is used to predict the training set $B$, and
   model $M_B$ is used to predict training set $A$ to obtain
   new pseudo surrounding boxes;
4. For each sample in $A$, if $IoU_{p\&o} < \gamma$, the predicted
   value of the sample is used as the new pseudo
   surrounding box. If $IoU_{p\&o} \geq \gamma$, the original pseudo
   surrounding box is used. If the fetal heart region is not
   identified, the original pseudo surrounding box is used.
   The updated pseudo surrounding box for $A$ is $A'$, and
   the same operation is performed with dataset $B$ to
   obtain $B'$. $IoU_{p\&o}$ represents the overlap between the
   predicted and the original pseudo surrounding box;
5. $A'$ and $B'$ are combined as a new training set $T'$ to
   train a new heart localization model $M_{temp}$;
6. Set $T = T'$ and repeat the process in steps 1-5
   multiple times to obtain the final model output $M_{out}$;
OUTPUT: The inferential cardiac localization model $M_{out}$.
```

Theoretically, ultrasound structures of fetal hearts with weak noise lead to better network feature extraction results and more accurate structural focus. Conversely, samples with high noise levels are more difficult to learn and fit. Suppose that the input to the regression model is $X$ and that the labeled box surrounding the target under accurate annotation conditions, that is, without noise, is represented as $Y_{gt}$. The pseudo- surrounding box generated by thresholding the class activation map is $Y_p$. We assume that $Y_p$ is obtained by adding noise $\varepsilon$ to $Y_{gt}$, that is, $Y_p = Y_{gt} + \varepsilon$. Since the noise $\varepsilon$ added to $Y_{gt}$ varies in intensity, we assume that $\varepsilon$ includes both strong noise $\varepsilon_s$ and weak noise $\varepsilon_w$. Therefore, if the training data $X$ contains samples with both strong and weak noise, the mathematical expectation of the noise should be calculated as follows:

$$E\left(\varepsilon_w\right) < E\left(\varepsilon\right) < E\left(\varepsilon_s\right) \qquad (1)$$

If the predicted pseudo surrounding box deviates only slightly from the original pseudo surrounding box, the sample is likely a weak noise sample, and its label should not be updated. However, if there is a significant deviation between the predicted and original surrounding boxes, the sample is likely a strong noise sample, and its label should be updated. Since $E(\varepsilon) < E(\varepsilon_s)$, updating the labels should reduce the overall noise in the dataset. By training with dataset $A$ and making predictions based on dataset $B$, the noise level in dataset $B$ can be reduced. Similarly, training with dataset $B$ and making predictions based on dataset $A$ can reduce the noise level in dataset $A$. When datasets $A$ and $B$ are combined, the overall noise level can be reduced. Therefore, during the next iteration, training is performed at a lower noise level, and the noise level should converge to the lower limit of $\varepsilon_w$.

Therefore, we propose a multiview classification and localization network by introducing the MESA mechanism into view classification to simultaneously recognize multiple views of fetal heart ultrasound images and extract key structural regions. In this approach, the structural focus and recognition accuracy of view recognition are first improved. Then, the surrounding boxes can be generated based on the threshold selection and multiple iterative regression strategy to identify anatomical structure regions of interest for quality scoring. Furthermore, due to the nature of CNNs, they are usually black boxes. Thus, CNNs are very unfriendly from a clinical perspective. With the embedding of the MESA mechanism, we can highlight the feature map of view classification through the activated feature map, making the view classification of the fetal heart ultrasound images interpretable. The details of our proposed MCLN are depicted in Fig. 2.

### B. Image Quality Scoring

Contrastive learning is applicable to the task of similarity measurement to determine the distance between augmented pairs. In general, the fundamental framework of contrastive learning involves selecting a data sample known as an ‘anchor’, a data point from the same distribution as the anchor, referred to as a ‘positive’ sample, and another data point from a different distribution, referred to as a ‘negative’ sample [40]. The goal of the contrastive learning model is to minimize the distance between the anchor and positive samples in latent space while maximizing the distance between the anchor and negative samples. Since there are no quantitative indicators, the relationship between positive and negative samples must be defined. Therefore, we borrow the idea of contrastive learning and propose an improved contrastive learning network to deal with quantitative score learning in quality assessment of fetal heart ultrasound images. The feature distance between samples with small quality score differences is also close, and the feature distance between samples with large quality score differences should be larger. Here, we changed the contrast learning strategy to use image pairs and their corresponding quality scores to compose input pairs instead of simple data augmentation. In addition to the original absolute loss function, we propose the feature relative loss function and the predicted value relative loss function to optimize the network. The relative feature loss measures the feature differences between samples to approximate the score difference between samples. This avoids the error of the absolute loss, and instead of directly calculating the similarity measure between two features to determine whether they are matched, the relative quality score is used for regression calculation. The predicted value relative loss represents the fact that the relative difference between the predicted value of two samples and the relative difference between the ground truth value of the two samples are consistent; otherwise, there is loss. This can be interpreted as a ranking loss for two samples. The sign of the score difference indicates the order of the quality scores for the two images. The network considers not only the absolute loss but also the relative loss of the input sample pairs. When the absolute loss is difficult to optimize, the distribution of the sample features can be further constrained by the relative loss.

We suppose that the input image pair is $x_i$ and $x_j$, the corresponding quality scores are $y_i$ and $y_j$ and the convolutional neural network with shared parameters can extract features $G_W(x_i)$ and $G_W(x_j)$, respectively. The network learns the mapping relationship of features to scores, so it passes the features through the fully connected (FC) layer. Here, it is assumed that the FC mapping of the predicted absolute score is $Fc_{abs}(x)$, the FC mapping of the predicted relative score is $Fc_{rela}(x)$, and the predicted absolute quality scores of images $x_i$ and $x_j$ are $y'_i$ and $y'_j$, respectively. This can be expressed as:

$$y'_i = Fc_{abs}\ (G_W\ (x_i)) \tag{2}$$

Here, the relative features of images $x_i$ and $x_j$ are $G_W(x_i)$ -$G_W(x_j)$, and the predicted relative score is:

$$y'_{rela(i,j)} = Fc_{rela}\ (G_W\ (x_i) - G_W\ (x_j)) \tag{3}$$

In the training process, the ICLN combines absolute loss $L_{abs}$, feature relative loss $L_{rela-f}$, and predicted value relative loss $L_{rela-p}$. $L_{rela-f}$ encourages the feature differences between samples to approximate the score difference between samples, which considers the relationship between samples. $L_{rela-f}$ can be seen as a ranking loss for two samples, because the sign of the score difference indicates the score order of two images. $L_{rela-f}$ considers both the magnitude and sign of the score difference.

$$L_{abs} = \frac{1}{N} \sum_i |y'_i - y_i| \tag{4}$$

$$L_{rela-f} = \frac{1}{N_p} \sum_{(i,j)} \left| y'_{rela(i,j)} - (y_i - y_j) \right| \tag{5}$$

$$L_{rela-p} = \frac{1}{N_p} \sum_{(i,j)} \left| (y'_i - y'_j) - (y_i - y_j) \right| \tag{6}$$

where $N$ is the batch size, $N_p$ denotes the number of paired samples $(i,\ j)$, so that the total loss function $L$ becomes

$$L\ = L_{abs}\ + \lambda_1 L_{rela-f} + \lambda_2 L_{rela-p} \tag{7}$$

where $\lambda_1 > 0$ and $\lambda_2 > 0$ are tradeoff hyperparameters that control the relative importance of the three terms.

## IV. Experimental Setup

### A. Dataset

We collected data from pregnant women who underwent B-mode fetal echocardiography examinations between 2018 and 2021. Our dataset consisted of 3596 2D ultrasound data samples obtained at a single center and 2421 2D ultrasound data samples obtained at multiple centers. For each scan we had access to freeze-frame images saved by the sonographers during the exam. The standard view was obtained according to the guidelines, and the specific schematic diagram is shown in Fig. 1. Furthermore, we established a category labeled as “other” for non-cardiac views, such as arms, hands, feet, bladder, diaphragm, coronal face and placenta views. Overall, our dataset consists of 60,526 images of standard views and 10,400 images of "other" views, including 41,899 standard view images obtained from a single center and 18,627 standard view images obtained from multiple centers, as well as 7,200 "other" view images from a single center and 3,200 "other" view images obtained from multiple centers. To train the components in the MCLN, up to 47,699 images from 3,596 cases were used. All training sets were acquired from the Key Laboratory of Maternal Fetal Medical Research and provided by Anzhen Hospital. Two datasets independent from the training dataset were used to evaluate the MCLN model performance. The two test sets were 1,400 single-center test sets and 21,827 multicenter test sets. The single-center data came from Anzhen Hospital and the multicenter data came from 38 medical institutions in the 13th Five-Year National Key Research and Development Plan (2018YFC1002300). The study protocols and procedures followed the protocols of the Declaration of Helsinki and were approved by the ethics committee of Beijing Anzhen Hospital (Approval No.2019030). The data were obtained from pregnant women aged 17–47 at a gestational age of 17–40 weeks. The median age at pregnancy was 29.8 $\pm$ 4.14 years, while the median gestational age was 27.43 $\pm$ 3.88 weeks.

For the datasets in the ICLN, we randomly selected 1,000 cases for each slice for expert annotation, with a total of 6000 images, 80% of which were used as the training set, and the remaining 20% were used as the test set. Fivefold cross-validation was conducted. The total score for each view is 10 points, and the assessment criteria for six planes has been validated by experts for many rounds. Here, we developed a complete clinical evaluation system for each view of the fetal heart that considers various factors such as the anatomical structure of fetal heart ultrasound, image clarity, and parameter settings.

For data annotation, we independently developed an annotation system to annotate the view classification and quality scoring results. The system has been well-designed to take into account the standardization and systematization of annotations. A double confirmation mechanism is adopted, first completed by senior echocardiographers, and then further verified by certified cardiologists. The process design includes data upload, grouping, annotation, review, approval, rejection, etc., achieving standardized and efficient completion of annotations. Therefore, the actual annotation process is a rapid process, where for view classification, each expert only needs to select the label for the data, which takes approximately 10 seconds per image. Quality evaluation only requires scoring each image, which takes an average of about 2–3 minutes per case.

## B. Training

The experiments in this article comprise two parts: MCLN and ICLN. To train the MCLN, we deployed a relatively high momentum of 0.9. The learning rate was set to 0.0002, with a gradual decrease of 0.5 every 1000 iterations. The batch size was set to 64, and each image was resized to 224∗224 before being input into the network. Moreover, to select more accurate bounding boxes containing the heart structure, we selected an optimal threshold and performed multiple iterations according to the Algorithm 1. The optimal number of iterations was selected to identify the best bounding box for quality assessment task in the ICLN.

For the ICLN, we used random sample pairs to conduct contrastive learning network training and learn the absolute and relative differences between samples. We also considered the sampling balance in the ICLN. Since the actual data scores are distributed according to a normal distribution, there are fewer samples with particularly high and low scores and more samples with intermediate scores. Therefore, during sampling, we can assign larger sampling probabilities to samples with small sample numbers and smaller sampling probabilities to samples with large sample numbers to balance the quality of sample pairs. Furthermore, data augmentation strategies were applied to samples with small sample sizes. To maintain the quality of the images, only left-right flipping and rotation operations were utilized for data augmentation. We trained the ICLN with different loss functions, which are shown in (4), (5) and (6).

All the reported results are from our implementation that used the Pytorch framework [41] and Python running on an Nvidia v100. The learning rate was decayed with a factor of 0.5 when the training loss did not decrease within 5 consecutive epochs. If the MAE based on the validation set did not decrease within 20 consecutive epochs, training was stopped.

## C. Evaluation Metric

We selected the precision, recall and F1 score as the main evaluation metrics for view classification, and we used the average F1 score over different classes because this approach increases the sensitivity to imbalances between classes. The F1 score is calculated as follows:

$$F1 \text{ score } = \frac{2R\text{ecall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \tag{8}$$

where recall is the ratio of correctly predicted positive observations to all observations in the actual class and precision is the ratio of correctly predicted positive observations to the total number of predicted positive observations. In Section V, we used the F1 score to measure the performance for view recognition.

We have plotted a confusion matrix for the multi-classification results, which allows us to clearly see the accuracy of view classification. For the ICLN, we selected the PCC and MSE as the main evaluation metrics. The PCC is a variable used to measure the "degree of linear correlation" of two variables, which is defined as the quotient of the product of the covariance of the two variables and the standard deviation of the two variables.

Typically, the total covariance and the standard deviation of the variables are difficult to obtain, and the sample covariance and the sample standard deviation of the variables are used for alternative estimation. PCC values range from $-1$ to 1, with a negative sign indicating a negative correlation, a positive sign indicating a positive correlation, an absolute value closer to 1 indicating a stronger linear correlation, and a PCC value of 0 indicating no correlation between two variables.

When PCC is used to measure the quality assessment algorithm, usually, the closer to 1 the PCC is, the better the quality of the assessment algorithm.

We assume that the two variables are $X, Y$, and the observed values of the sample are $x_1, x_2, \ldots, x_N$ and $y_1, y_2, \ldots, y_N$,

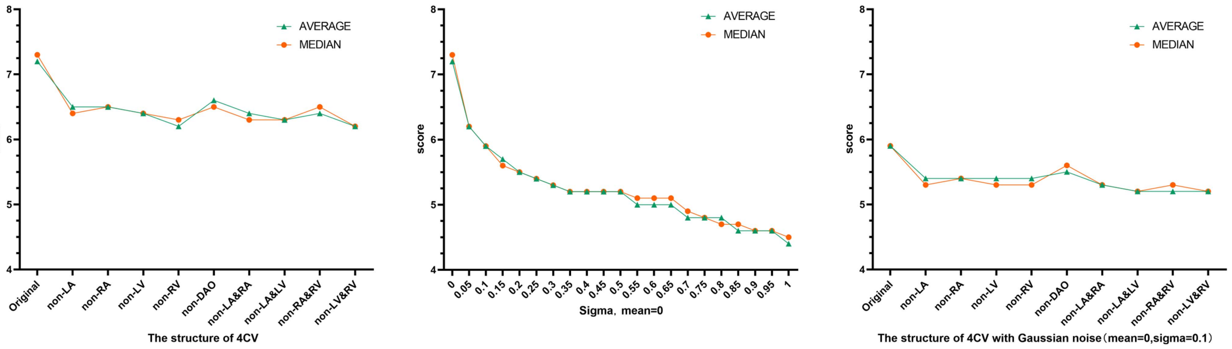

Fig. 3. Quality score changes with added Gaussian noise and missing structures. The structure includes the left atrium (LV), left atrium (LA), right ventricle (RV), right atrium (RA), and descending aorta (DAO).

where N is the sample size; then, the PCC calculation formula can be expressed as follows:

$$PCC = \frac{N \times \sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{\sqrt{N \times \sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2} \sqrt{N \times \sum_{i=1}^{N} y_i^2 - \left(\sum_{i=1}^{N} y_i\right)^2}} \quad (9)$$

The MSE represents the average of the squared difference between two variables. Assuming that the two variables are $X, Y$ and the observed values of the sample are $x_1, x_2, \ldots, x_N$ and $y_1, y_2, \ldots, y_N$, where N is the sample size, the calculation formula for the MAE can be expressed as follows:

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (x_i - y_i)^2 \quad (10)$$

For the purpose of verifying the effect of the ICLN in clinical practice, we used the proportion of the absolute errors between the predicted value and the ground truth value less than 1 point, less than 2 points, and less than 3 points. Here, let the number of samples in the test set be N, the ground truth quality scores be $y_1, y_2, \ldots, y_N$, and the quality scores predicted by the model be $y_1', y_2', \ldots, y_N'$. $P_{\varepsilon \le 1}$, $P_{\varepsilon \le 2}$, and $P_{\varepsilon \le 3}$ represent the proportion of absolute errors between the predicted value and the ground truth value less than 1 point, less than 2 points, and less than 3 points, respectively. The expression of each measurement index can be denoted as follows:

$$P_{\varepsilon \le 1} = \frac{1}{N} \sum_{i=1}^{N} 1_{|y_i - y_i'| \le 1} \quad (11)$$

$$P_{\varepsilon \le 2} = \frac{1}{N} \sum_{i=1}^{N} 1_{|y_i - y_i'| \le 2} \quad (12)$$

$$P_{\varepsilon \le 3} = \frac{1}{N} \sum_{i=1}^{N} 1_{|y_i - y_i'| \le 3} \quad (13)$$

## V. Results and Analysis

### A. Integrity Verification of the Proposed Network

To verify the effectiveness of our proposed method for structural integrity and image quality and clarity, we conducted experiments by occluding anatomical structures and adding Gaussian noise. As illustrated in Fig. 3, we selected 4CV for the experiment, in which the anatomical structure included the left ventricle (LV), left atrium (LA), right ventricle (RV), right atrium (RA), and descending aorta (DAO). We increased the intensity of Gaussian noise by changing the value of sigma, which ranged from 0 to 1. Fig. 3 shows that structural occlusions and decreased quality score are positively correlated; that is, more structural occlusions lead to lower quality scores. Furthermore, the loss of the ventricle is even more pronounced than the loss of the atrium. This may be due to the relatively large size of the ventricle, which has a greater impact on the quality score. For Gaussian noise, we see that the image quality score gradually decreases with increasing noise value. The maximum decrease is at the node where the noise is first added. We see that the quality score of the descending aorta occlusion decreases less than in the other structures, probably because this structure is a relatively small part of the overall structure of the fetal heart. For the combination of Gaussian noise and missing structure, we used Gaussian noise with sigma $= 0.1$, and the quality score of the combination was lower than that of missing structures and adding noise alone.

### B. Verification of Multiview Recognition

Experimental validation of multiview recognition was performed on single-center and multicenter datasets. Here, precision, recall, and the F1 score were used to evaluate the view classification accuracy. We conducted experiments based on the MCLN model to verify the improvement in classification accuracy with seven classes, several typical CNN models, including AlexNet [42], DenseNet [43], ResNet [44] and specifically designed for echocardiograms such as, SonoNet [27] and SEVDR [21] were reimplemented for comparison. As shown in Fig. 4, we observe that the proposed network showed the best performance in all seven categories and was followed by DenseNet, SEVDR, ResNet, and SonoNet, with AlexNet performing the worst in multiview recognition. This is attributed to the MESA attention mechanism better capturing information in multiple subspaces during feature extraction and obtaining more global information within the attention range. Therefore, this approach is more sensitive to the relatively fixed structural relationship of various views of fetal heart ultrasound images. These improvements make the classification results more focused on the structural information of the fetal heart rather than noise

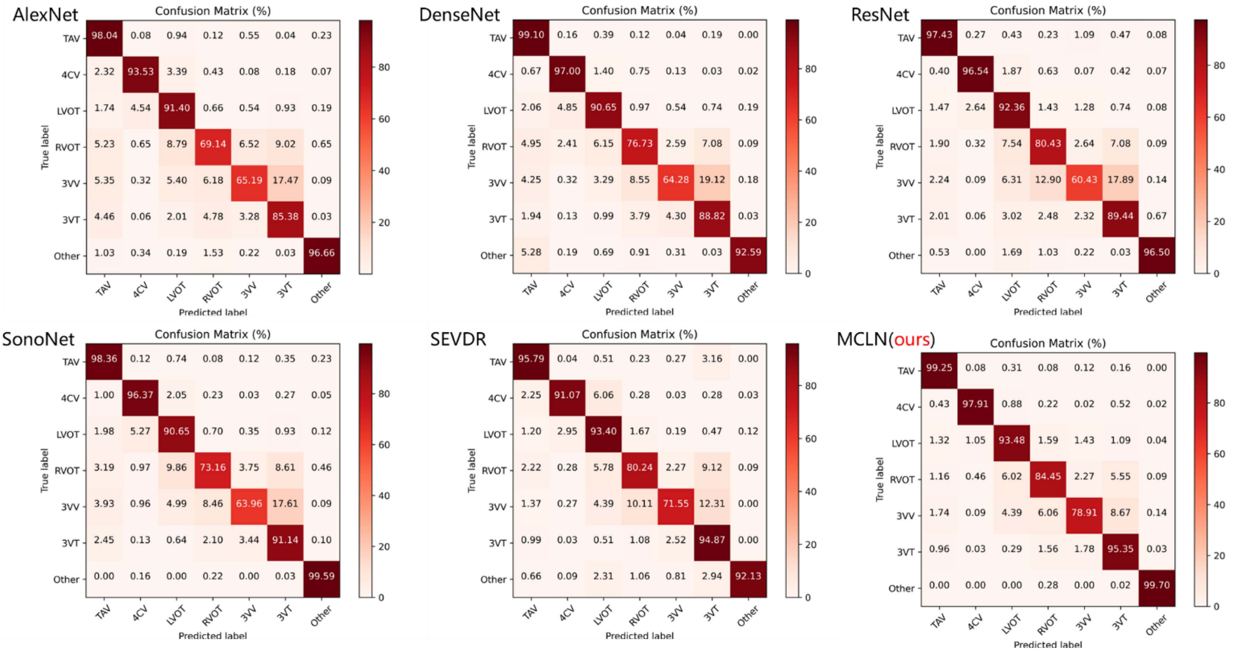

Fig. 4. Confusion matrix between different fetal echocardiographic views from different methods based on multicenter data.

and other interference factors. For data from a single center, the normalization process is effective, resulting in high image classification accuracy. However, for the multi-center classification results, our network improved significantly, especially for the right ventricular outflow tract view (RVOT), three-vessel view (3VV) and three-vessel tracheal view (3VT), of which the RVOT improved by 4.53∼11.33% in terms of F1-score, the 3VV improved by 4.98∼13.61% in terms of F1-score. The 3VT was improved by 5.09∼10.08% in terms of F1-score. Compared to the model performance for TAV, 4CV, LVOT, and other view, the performance of all models for RVOT, 3VV, and 3VT has decreased. The most obvious decline was in RVOT, which may be due to its inclusion of two kinds of views, and the large difference in data acquisition of multiple centers, which increases the difficulty of identification. From the Fig. 4, we can also observe that the 3VV and RVOT view were easily classified as 3VT, the 3VV also was easy to get misclassified as RVOT. These misclassifications mainly occur because these views are relatively close to each other. In addition, during the scanning process, the fetal heart moves, and the image slices capture transitions between views, making misclassification more likely. Through further analysis of the misclassified images, we found that most of the misclassifications in RVOT are between RVOT-1,3VV and 3VT images. Moreover, the training dataset for the 3VV section was small, which may be the reason for the generally low recognition accuracy of the 3VV section. In the future, we will add more training data to improve the performance of the proposed model.

## C. Verification of Anatomical Structure Localization

To validate whether each view is applicable for clinical feature classification and verify the effectiveness of the MESA mechanism in classifying each view of the fetal heart, we performed gradient-weighted class activation mapping (Grad-CAM) experiments [45]. We used the features of the penultimate convolutional layer of the classification network of each view of the fetal heart. Both experiments show that the view classifier makes its decisions based on clinically relevant image features, and the MESA mechanism makes the classification more focused on the anatomical structure. Fig. 5 shows the Grad-CAM maps of the images based on original ResNet model and the model with the embedded MESA mechanism network. The figure shows that the embedded MESA mechanism causes the model to focus more on key anatomical structures, and the scope of the attention area is more accurate. In addition, the weight outside the anatomical structure is lower. For example, in the first column, ResNet focused on a larger area, even including the area outside the thoracic cavity, while the network with the embedded MESA mechanism focused more on the stomach bubble, umbilical vein and other key anatomical structures, which was also verified in other views. This is because the attention mechanism extracts the important information from the global features and ignores interference information such as noise and background, which makes the classification network pay more attention to the most important regions and improves the classification performance.

## D. Verification Results of the Loss Function

Table I shows the results of each evaluation metric of 4CV of the fetal heart. Among them, ICLN with $L_{\text{abs}}$, $L_{\text{rela-f}}$ and $L_{\text{rela-p}}$ showed the best performance, followed by ICLN with $L_{\text{abs}}$ and $L_{\text{rela-f}}$ and ICLN (only $L_{\text{abs}}$), and the single network had the worst performance. The 4CV showed that more than 97% of the samples having been predicted and ground truth values within 2 points, and the mean absolute error was less than 1 point, indicating the effectiveness of the ICLN. This may be because the anatomical structure of the four-chamber

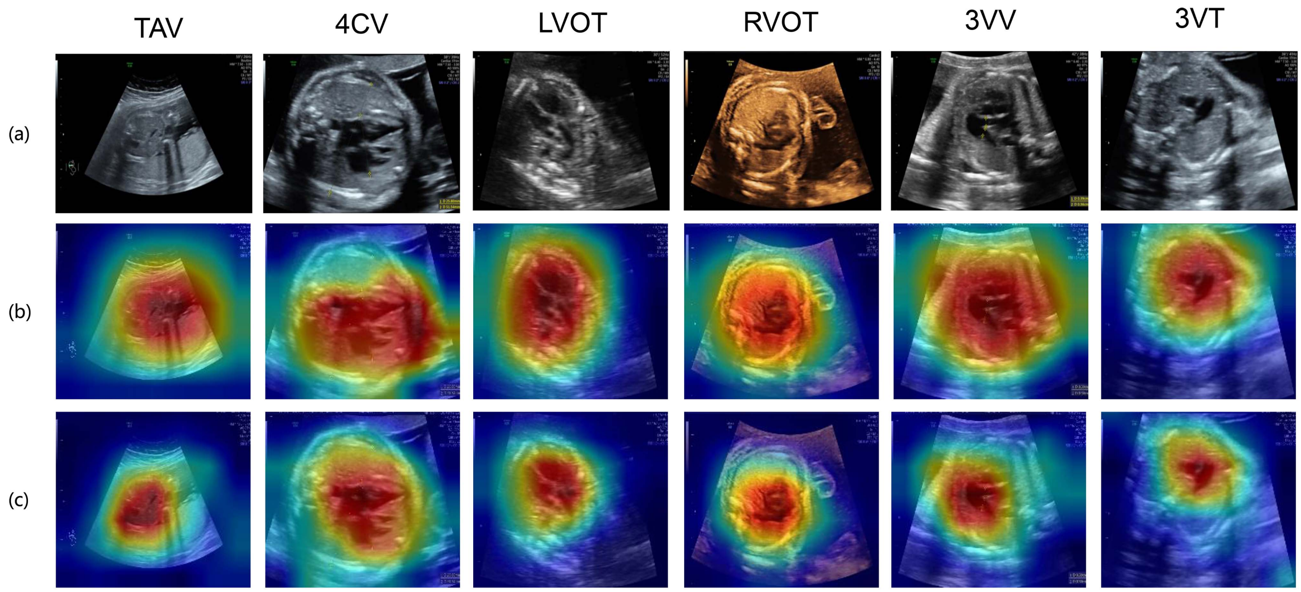

Fig. 5. Comparison of the Grad-CAM map between different methods, (a) is raw data of six views. (b) Is the original ResNet [42], and (c) is the ResNet incorporating the enhanced multihead self-attention mechanism.

TABLE I
Comparison of Different Indexes of Different Loss Functions in Multiple Views

| View | Method | PCC | MSE | $P_{\varepsilon\leq1}$ | $P_{\varepsilon\leq2}$ | $P_{\varepsilon\leq3}$ |
|---|---|---|---|---|---|---|
| 4CV | Single Network | 0.673 | 1.563 | 0.451 | 0.821 | 0.955 |
| | ICLN (only $L_{\text{abs}}$) | 0.679 | 1.447 | 0.438 | 0.864 | 0.964 |
| | ICLN ($L_{\text{abs}}$ *and* $L_{\text{rela-f}}$) | 0.764 | 1.018 | 0.695 | 0.961 | 0.990 |
| | **ICLN ($\boldsymbol{L_{\text{abs}}}$ , $\boldsymbol{L_{\text{rela-f}}}$ , $\boldsymbol{L_{\text{rela-p}}}$)** | **0.780** | **0.979** | **0.705** | **0.971** | **0.997** |

Bold values indicate best performance.

heart view is numerous and easy to identify, with relatively obvious characteristics. The expert annotation also increased the difficulty, which had an impact on the accuracy of the ICLN. In addition, the model using the contrastive network achieved better results than the model using the single network because the contrastive network takes the sample pair composed of the two samples as the input during training, which can optimize the relative error as well as the absolute error, which is equivalent to introducing the human relative aesthetic mechanism. Moreover, the feature relative loss and predicted value relative loss were increased to further improve the performance of the ICLN.

## E. Comparison With Deep Learning Algorithms

We selected state-of-the-art image quality assessment networks and a comparative learning network for comparison to verify the effectiveness of the ICLN. Table II provides a comparison between the proposed network, typical deep learning networks and specifically designed netoworks for echocardiograms (IQA-CNN [28], DeepIQA [46], B-CNN [47], CMC [48], SupCon [49], AES [22], D-CNN [23]).The results show that the highest PCC and lowest MSE were obtained using the ICLN model with $L_{\text{abs}}, L_{\text{rela}-\text{f}}$ and $L_{\text{rela}-\text{p}}$; that is, the best quality assessment performance was obtained with the fetal heart ultrasound images. In addition, our proposed network had higher values than the other networks in terms of the proportion of absolute errors between the predicted value and the ground truth value for all views of fetal heart ultrasound images, and the ICLN model with $L_{\text{abs}}, L_{\text{rela}-\text{f}}$ and $L_{\text{rela}-\text{p}}$ performed better than the other networks. The $P_{\varepsilon\leq1}$ of our proposed method was better than that of the other methods in 4CV, reaching 9.1%–29%. In addition, the lowest PCC and highest MSE for the 3VT view was obtained using SupCon, while for 4CV, the performance rankings of the other networks were, in sequence from best to worst performance, B-CNN IQA-CNN, AES, CMC, D-CNN, DeepIQA and SupCon.

## VI. Discussion and Conclusion

We proposed a framework comprising two CNN-based networks (MCLN and ICLN), which were used to perform multiview recognition and quality assessment, respectively. Experiments on a multiview fetal cardiac ultrasound dataset demonstrated the effectiveness of the proposed framework. In addition, a comparison with other state-of-the-art deep learning networks demonstrated the generalization and adaptability of the proposed framework. The proposed MCLN outperformed other state-of-the-art networks in fetal heart six standard view recognition. To our knowledge, this is the first time that six views of the fetal heart have been classified and interpretable effects have been achieved based on attention mechanisms. This has great potential for the study of other fetal organs.

Actually, the proposed ICLN addresses the problem of how to automatically select the most standard view from a fetal heart

TABLE II
COMPARISON OF DIFFERENT METRICS BETWEEN OUR PROPOSED NETWORK AND OTHER STATE-OF-THE-ART NETWORKS IN MULTIPLE VIEWS

| View | Method | PCC | MSE | $P_{\varepsilon\leq1}$ | $P_{\varepsilon\leq2}$ | $P_{\varepsilon\leq3}$ |
|---|---|---|---|---|---|---|
| 4CV | IQA-CNN[28] | 0.594 | 1.246 | 0.588 | 0.909 | 0.974 |
| | DeepIQA[46] | 0.536 | 1.457 | 0.419 | 0.844 | 0.977 |
| | B-CNN[47] | 0.638 | 1.197 | 0.610 | 0.929 | 0.981 |
| | CMC[48] | 0.574 | 1.397 | 0.535 | 0.844 | 0.932 |
| | SupCon[49] | 0.355 | 1.785 | 0.497 | 0.769 | 0.886 |
| | AES[22] | 0.588 | 1.386 | 0.614 | 0.912 | 0.966 |
| | D-CNN[23] | 0.566 | 1.430 | 0.574 | 0.877 | 0.954 |
| | **ICLN** | **0.780** | **0.979** | **0.705** | **0.971** | **0.997** |
| 3VT | IQA-CNN[28] | 0.166 | 2.200 | 0.304 | 0.572 | 0.793 |
| | DeepIQA[46] | 0.436 | 2.018 | 0.386 | 0.693 | 0.864 |
| | B-CNN[47] | 0.338 | 2.119 | 0.318 | 0.618 | 0.836 |
| | CMC[48] | 0.140 | 2.715 | 0.346 | 0.536 | 0.704 |
| | SupCon[49] | 0.107 | 2.887 | 0.361 | 0.521 | 0.639 |
| | AES[22] | 0.421 | 2.306 | 0.375 | 0.663 | 0.830 |
| | D-CNN[23] | 0.386 | 2.620 | 0.322 | 0.583 | 0.819 |
| | **ICLN** | **0.584** | **1.811** | **0.454** | **0.707** | **0.897** |

Bold values indicate best performance.

scan to diagnose FCHD. This is meaningful for the clinical setting because the standard view appears quickly during fetal heart scanning, which makes it difficult for the sonographer to obtain the image manually; this increases the scan time and is not pleasant for the pregnant woman. Different from previous structure-based detection networks, we do not need to annotate each structure or substructure. By adding the relative score and the relative loss, we reduce the error of the image quality assessment. In addition, embedding MESA and proposed multiple iterative regression strategy in the first step are helpful for quality scoring and makes the classified views more accurately focus on the anatomical structure rather than the interference information, such as acoustic shadows and noncardiac structures. In addition, the ICLN considers both the standardization of anatomical structures and the quality clarity of images, achieving results comparable to those achieved by experts with little annotation expense, which has great implications for clinical application.

In the data collection process, we included six cardiac-related views and views of other organs examined by prenatal ultrasound examinations, such as the head, face and limbs, etc. In our dataset, we defined these views as “other”, and 7000 “other” cases were included in the training set, and 3200 “other” cases were included in the external test set. However, we cannot cover all non-cardiac views during the training process. Some scholars have carried out relevant research to address the uncertainty problem in view classification. For example, Gu et al. proposed that the Efficient-Evidential Network could provide uncertainty prediction based on input samples [50]. Therefore, clinicians can be prompted to conduct interactive operations such as recollecting data. Liao et al. modeled the intra-observer variability caused by the uncertainty of the annotated data and demonstrated the effectiveness of their model in ultrasound image quality assessment [51]. Moreover, the prerequisite is that all views have been collected in the process of fetal heart scanning, which has certain requirements for sonographers. Therefore, to fundamentally resolve the clinical problem of data acquisition of the fetal heart, it is essential to establish automatic navigation of fetal heart ultrasound scanning, that is, the automatic positioning of the fetal heart views and the transformation relationship between the views. In addition, the view recognition result and the image quality scores can be fed back to the navigation to guide the most standard view of the fetal heart. Thus, the problem of obtaining fetal cardiac ultrasound can be truly resolved, and a good data basis can be provided for the screening and diagnosis of FCHD. Meanwhile, the gestational age of the data in this article was 17–40 weeks, but there is a lack of data in the first trimester, such as 11–16 weeks. Extracting standard views of diagnostic quality in the first trimester is of great significance for the early detection of FCHD. In addition to view-level sectional recognition and quality assessment, disease-level image analysis, such as multiview automatic screening and automatic diagnosis of major FCHDs, is the direction of future work.

## REFERENCES

[1] K. M. Verdurmen et al., “A systematic review of prenatal screening for congenital heart disease by fetal electrocardiography,” *Int. J. Gynecol. Obstet.*, vol. 135, no. 2, pp. 129–134, Nov. 2016, doi: 10.1016/j.ijgo.2016.05.010.

[2] I. Germanakis and S. Sifakis, “The impact of fetal echocardiography on the prevalence of liveborn congenital heart disease,” *Pediatr. Cardiol.*, vol. 27, no. 4, pp. 465–472, Jul./Aug. 2006, doi: 10.1007/s00246-006-1291-6.

[3] GBD 2017 Congenital Heart Disease Collaborators, “Global, regional, and national burden of congenital heart disease, 1990-2017: A systematic analysis for the global burden of disease study 2017,” *Lancet Child Adolesc. Health*, vol. 4, no. 3, pp. 185–200, Mar. 2020, doi: 10.1016/s2352-4642(19)30402-x.

[4] Q. M. Zhao, F. Liu, L. Wu, X. J. Ma, C. Niu, and G. Y. Huang, “Prevalence of congenital heart disease at live birth in China,” *J. Pediatrics*, vol. 204, pp. 53–58, Jan. 2019, doi: 10.1016/j.jpeds.2018.08.040.

[5] X. Zhang et al., “The significance of an integrated management mode of prenatal diagnosis-postnatal treatment for critical congenital heart disease in newborns,” *Cardiovasc. Diagnosis Ther.*, vol. 11, no. 2, pp. 447–456, Apr. 2021, doi: 10.21037/cdt-20-892.

[6] N. Chitra and I. B. Vijayalakshmi, “Fetal echocardiography for early detection of congenital heart diseases,” *J. Echocardiogr.*, vol. 15, no. 1, pp. 13–17, Mar. 2017, doi: 10.1007/s12574-016-0308-2.

[7] N. J. Dudley and E. Chapman, “The importance of quality management in fetal measurement,” *Ultrasound Obstet. Gynecol.*, vol. 19, no. 2, pp. 190–196, Feb. 2002, doi: 10.1046/j.0960-7692.2001.00549.x.

[8] B. Rahmatullah, I. Sarris, A. Papageorghiou, and J. A. Noble, “Quality control of fetal ultrasound images: Detection of abdomen anatomical landmarks using AdaBoost,” in *Proc. IEEE Int. Symp. Biomed. Imag.: From Nano to Macro*, 2011, pp. 6–9.

[9] L. J. Salomon and Y. Ville, “Quality control of prenatal ultrasound,” *Ultrasound Rev. Obstet. Gynecol.*, vol. 5, no. 4, pp. 297–303, Dec. 2005, doi: 10.3109/14722240500415419.

[10] L. J. Salomon, J. P. Bernard, M. Duyme, B. Doris, N. Mas, and Y. Ville, “Feasibility and reproducibility of an image-scoring method for quality control of fetal biometry in the second trimester,” *Ultrasound Obstet. Gynecol.*, vol. 27, no. 1, pp. 34–40, Jan. 2006, doi: 10.1002/uog.2665.

[11] J. S. Carvalho et al., “ISUOG practice guidelines (updated): Sonographic screening examination of the fetal heart,” *Ultrasound Obstet. Gynecol.*, vol. 41, no. 3, pp. 348–359, Mar. 2013, doi: 10.1002/uog.12403.

[12] S. Wang et al., “Consensus on the medical model and technical process of multidisciplinary diagnosis and treatment and precision integrated prevention and management of fetal heart disease in maternal-fetal medicine (Part II): Consensus on detailed risk stratification diagnosis technology of fetal echocardiography,” *Chin. J. Perinatal Med.*, vol. 25, pp. 481–487, 2022.

[13] M. Itsukaichi et al., “Effectiveness of fetal cardiac screening for congenital heart disease using a combination of the four-chamber view and three-vessel view during the second trimester scan,” *J. Obstet. Gynecol. Res.*, vol. 44, no. 1, pp. 49–53, Jan. 2018, doi: 10.1111/jog.13472.

[14] G. S. Bak et al., “Detection of fetal cardiac anomalies: Cost-effectiveness of increased number of cardiac views,” *Ultrasound Obstet. Gynecol.*, vol. 55, no. 6, pp. 758–767, Jun. 2020, doi: 10.1002/uog.21977.

[15] J. S. Carvalho, E. Mavrides, E. A. Shinebourne, S. Campbell, and B. Thilaganathan, “Improving the effectiveness of routine prenatal screening for major congenital heart defects,” *Heart*, vol. 88, no. 4, pp. 387–391, Oct. 2002, doi: 10.1136/heart.88.4.387.

[16] A. Ghorbani et al., “Deep learning interpretation of echocardiograms,” *NPJ Digit. Med.*, vol. 3, Jan. 2020, Art. no. 10, doi: 10.1038/s41746-019-0216-8.

[17] P. Zhu and Z. Li, “Guideline-based learning for standard plane extraction in 3-D echocardiography,” *Proc. SPIE*, vol. 5, no. 4, Oct. 2018, Art. no. 044503, doi: 10.1117/1.jmi.5.4.044503.

[18] M. C. Hemmsen, T. Lange, A. H. Brandt, M. B. Nielsen, and J. A. Jensen, “A methodology for anatomic ultrasound image diagnostic quality assessment,” *IEEE Trans. Ultrason. Ferroelect. Freq. Control*, vol. 64, no. 1, pp. 206–217, Jan. 2017, doi: 10.1109/tuffc.2016.2639071.

[19] N. Van Woudenberg et al., “Quantitative echocardiography: Real-time quality estimation and view classification implemented on a mobile Android device,” in *Proc. Simul., Image Process., Ultrasound Syst. Assist. Diagnosis Navigation*, 2018, pp. 74–81.

[20] A. Narang et al., “Utility of a deep-learning algorithm to guide novices to acquire echocardiograms for limited diagnostic use,” *JAMA Cardiol.*, vol. 6, no. 6, pp. 624–632, Jun. 2021, doi: 10.1001/jamacardio.2021.0185.

[21] L. Wu et al., “Standard echocardiographic view recognition in diagnosis of congenital heart defects in children using deep learning based on knowledge distillation,” *Front. Pediatrics*, vol. 9, May 2021, Art. no. 770182, doi: 10.3389/fped.2021.770182.

[22] A. H. Abdi et al., “Automatic quality assessment of echocardiograms using convolutional neural networks: Feasibility on the apical four-chamber view,” *IEEE Trans. Med. Imag.*, vol. 36, no. 6, pp. 1221–1230, Sep. 2017, doi: 10.1109/tmi.2017.2690836. Erratum in: IEEE Trans Med Imaging. 2017 Sep;36(9):1992. PMID: 28391191.

[23] J. Dong et al., “A generic quality control framework for fetal ultrasound cardiac four-chamber planes,” *IEEE J. Biomed. Health Inform.*, vol. 24, no. 4, pp. 931–942, Apr. 2020, doi: 10.1109/jbhi.2019.2948316.

[24] L. Wu, J. Z. Cheng, S. Li, B. Lei, T. Wang, and D. Ni, “FUIQA: Fetal ultrasound image quality assessment with deep convolutional networks,” *IEEE Trans. Cybern.*, vol. 47, no. 5, pp. 1336–1349, May 2017, doi: 10.1109/tcyb.2017.2671898.

[25] H. Chen et al., “Ultrasound standard plane detection using a composite neural network framework,” *IEEE Trans. Cybern.*, vol. 47, no. 6, pp. 1576–1586, Jun. 2017, doi: 10.1109/tcyb.2017.2685080.

[26] Q. Yang et al., “Development of digital fetal heart models with virtual ultrasound function based on cardiovascular casting and computed tomography scan,” *Bioeng. (Basel)*, vol. 9, no. 10, Oct. 2022, Art. no. 524, doi: 10.3390/bioengineering9100524.

[27] CF Baumgartner et al., “SonoNet: Real-time detection and localisation of fetal standard scan planes in freehand ultrasound,” *IEEE Trans. Med. Imag.*, vol. 36, no. 11, pp. 2204–2215, Nov. 2017, doi: 10.1109/TMI.2017.2712367.

[28] L. Kang, P. Ye, Y. Li, and D. Doermann, “Convolutional neural networks for no-reference image quality assessment,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1733–1740.

[29] X. Liu, J. Van De Weijer, and A. D. Bagdanov, “RankIQA: Learning from rankings for no-reference image quality assessment,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1040–1049.

[30] V. Gattupalli, P. S. Chandakkar, and B. Li, “A computational approach to relative aesthetics,” in *Proc. IEEE 23rd Int. Conf. Pattern Recognit.*, 2016, pp. 2446–2451.

[31] A. Vaswani et al., “Attention is all you need,” in *Proc. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[32] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.

[33] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, “Self-attention generative adversarial networks,” in *Proc. Int. Conf. Mach. Learn.,* 2018, pp. 7354–7363.

[34] Y. Zhang, Y. Gong, H. Zhu, X. Bai, and W. Tang, “Multi-head enhanced self-attention network for novelty detection,” *Pattern Recognit.*, vol. 107, Nov. 2020, Art. no. 107486, doi: 10.1016/j.patcog.2020.107486.

[35] W. Wu, H. Chang, Y. Zheng, Z. Li, Z. Chen, and Z. Zhang, “Contrastive learning-based robust object detection under smoky conditions,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 4294–4301.

[36] X. Wang et al., “Transformer-based unsupervised contrastive learning for histopathological image classification,” *Med. Image Anal.*, vol. 81, 2022, Art. no. 102559, doi: 10.1016/j.media.2022.102559.

[37] J Sun et al., “Classification for thyroid nodule using ViT with contrastive learning in ultrasound images,” *Comput. Biol. Med.*, vol. 152, Jan. 2023, Art. no. 106444, doi: 10.1016/j.compbiomed.2022.106444.

[38] M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 818–833.

[39] C L Zhang, Y H Cao, and J. Wu, “Rethinking the route towards weakly supervised object localization,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 13457–13466.

[40] A. Jaiswal et al., “A survey on contrastive self-supervised learning,” *Technologies*, vol. 9, no. 1, 2021, Art. no. 2.

[41] A. Paszke et al., “Pytorch: An imperative style, high-performance deep learning library,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, vol. 32, pp. 8024–8035.

[42] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, vol. 25, pp. 1097–1105.

[43] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *Proc IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2261–2269.

[44] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[45] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.

[46] S. Bosse, D. Maniry, K. R. Müller, T. Wiegand, and W. Samek, “Deep neural networks for no-reference and full-reference image quality assessment,” *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 206–219, Jan. 2018, doi: 10.1109/TIP.2017.2760518.

[47] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang, “Blind image quality assessment using a deep bilinear convolutional neural network,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 1, pp. 36–47, Jan. 2020, doi: 10.1109/TCSVT.2018.2886771.

[48] Y. Tian, D. Krishnan, and P. Isola, “Contrastive multiview coding,” in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 776–794.

[49] P. Khosla et al., “Supervised contrastive learning,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 18661–18673.

[50] A. Gu et al., “Efficient echocardiogram view classification with sampling-free uncertainty estimation,” in *Proc. Int. Workshop Adv. Simplifying Med. Ultrasound*, 2021, pp. 139–148.

[51] 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. Imag.*, vol. 39, no. 6, pp. 1868–1883, Jun. 2020.