ISSN 2616-5775 Vol. 8, Issue 11: 1-13, DOI: 10.25236/AJCIS.2025.081101

Contrastive Cross-View Representation Learning for Echocardiographic View Classification

Jiawei Han[1,a], Xuande Zhang[1,b], Long Xu[2,c], Kunjing Pang[3,d], Xin Huang[2,e,*]

[1]School of Electronic Information and Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an, Shaanxi, China
[2]Faculty of Electrical Engineering and Computer Science, Ningbo University, Ningbo, Zhejiang, China
[3]Department of Echocardiography, Fuwai Hospital, National Center for Cardiovascular Diseases, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China
[a]Hanjiawei0919@163.com, [b]love_truth@126.com, [c]xulong1@nbu.edu.cn,
[d]pangkunjing2020@126.com, [e]huangxin@nbu.edu.cn
** Corresponding author*

Abstract: *The automatic analysis of echocardiography is of great significance in the diagnosis of cardiovascular diseases, and accurate view classification is the basis for achieving the automatic analysis of echocardiography. However, the existing automation methods are confronted with numerous challenges: they rely heavily on expensive human annotations, have insufficient generalization ability among people and devices, and fail to fully utilize the inherent anatomical consistency among different standard views. To address these limitations, this study proposes a novel structure-aware self-supervised learning framework. The core idea is to guide the model to learn the view-invariant representation by constructing positive sample pairs between different echocardiogram views of the same patient, thus eliminating the need for manual annotation. This method enables the model to effectively capture anatomical consistency across views, providing more robust features for downstream tasks. The experimental results show that the viewpoint classification performance of this method on four standard echocardiogram datasets has been significantly improved, effectively verifying its effectiveness and clinical application potential.*

Keywords: *Echocardiography, View Classification, Self-Supervised Learning, Structural Representation*

1. Introduction

Cardiovascular diseases (CVD) remain the leading cause of morbidity and mortality worldwide. According to reports from the World Health Organization (WHO) and the American Heart Association (AHA), more than 17 million people die from cardiovascular diseases each year, accounting for over 30% of global deaths[1]. In clinical practice, echocardiography has become the most widely used cardiac imaging method due to its advantages such as non-invusiveness, strong real-time performance and relatively low cost. It is routinely used for cardiac function assessment, structural abnormality diagnosis and disease follow-up[2,3].

In echocardiographic analysis, view classification is the foundation of all downstream tasks. Doctors usually obtain multiple standard views such as the parapical long axis (PLAX), apical four-compartment (A4C), apical two-compartment (A2C), and apical three-compartment (A3C) to comprehensively assess the structure and function of the heart[4]. Therefore, the accuracy of automatic view classification directly determines the reliability of subsequent analyses, including left ventricular function assessment, valve disease detection, and cardiomyopathy diagnosis. Any misclassification at this stage may spread the error to the subsequent diagnostic process, thereby leading to serious clinical risks. Therefore, ensuring high-precision view classification is an important prerequisite for intelligent echocardiography analysis.

In recent years, deep learning methods have been widely applied in the classification of echocardiographic images and have achieved remarkable progress. For instance, Madani et al. [5]demonstrated in their early work a model based on convolutional neural networks (CNNS) that could automatically classify 15 standard views with an accuracy rate close to 98%. Østvik et al.[6]further introduced real-time classification, achieving efficient clinical deployment. Meanwhile, the adoption of

ISSN 2616-5775 Vol. 8, Issue 11: 1-13, DOI: 10.25236/AJCIS.2025.081101

large-scale datasets has also accelerated technological progress; For instance, Naser et al.[7]trained CNNS on data from over 900 patients and consistently maintained an accuracy rate of over 96% in various test scenarios. In addition, Huang et al.[8]introduced a capsule network heuristic interpretable mechanism to enhance the clinical interpretability of the prediction results, while Kususnose et al.[9]verified the feasibility of the CNN-based model in clinical practice through five standard viewpoints. Some recent studies have been conducted on specific populations and diagnostic Settings: Wu et al.[10] trained models on large pediatric congenital heart disease datasets to emphasize the significant role of view recognition in pediatric cardiology; Another study extended the view classification to valve localization, making functional and hemodynamic analysis possible [11].

Despite a series of advancements, the existing methods still face numerous challenges. These methods rely heavily on large-scale, high-quality labeled data, and the acquisition cost of such labeled data is extremely high. Furthermore, their generalization ability is limited under equipment, patient populations or low-quality imaging conditions, often leading to performance degradation [5]. Meanwhile, most studies have focused on disease prediction or quantitative analysis rather than specifically optimizing fundamental view classification tasks, and there is still room for improvement in this field[12].

Against this backdrop, self-supervised learning (SSL) has emerged as a promising research paradigm in the field of computer vision, leveraging proxy tasks to fully exploit the potential of unlabeled data [13,14]. Contrastive learning methods, such as SimCLR[15] and MoCo [16], significantly improve the quality of representation by constructing positive and negative samples. Subsequent methods, such as BYOL [17] and SimSiam [18], eliminate the reliance on negative samples, thereby making the training process more stable and efficient. In terms of medical imaging, studies such as Models Genesis [19] and Med3D [20] have verified the effectiveness of SSL in CT and MRI tasks. However, in the application of echocardiography, its development remains relatively limited, especially in the classification of views, where the structural consistency of cross-views has not been fully utilized.

To address these limitations, we propose a structure-aware self-supervised learning method. The core idea is to guide the model to capture the structural invariance across views by constructing positive pairs among multiple views of the same patient. Specifically, we adopt Convolutional Network with Next-Generation Architecture (ConvNeXt) as the backbone encoder and introduce a structural information input construction strategy during the pre-training process to enhance the model's ability to represent the anatomical structure of the heart. In downstream classification tasks, compared with traditional methods, the accuracy of this method has increased by approximately 6-8%, highlighting its potential in clinical applications.

The structure of this article is arranged as follows: Section Two provides a review of the research related to the classification of echocardiographic images; Section Three presents a structure-aware self-supervised learning method; The fourth section reports on the experimental design, comparative analysis and its results. Finally, the fifth part summarizes this research and explores future research directions.

2. Materials

2.1 The Application of Deep Learning Methods in the Classification of Echocardiographic Views

Echocardiographic view classification, as an important basic task in clinical diagnosis, has attracted extensive research attention in recent years with the rapid development of deep learning methods. In a representative study first proposed by Madani et al. [5], a classification model based on CNN was developed, which could effectively distinguish 15 standard views, including 12 video-based views and 3 static views. Their method achieved an accuracy rate of 97.8% in video classification and 91.7% in static image classification, significantly outperforming the average accuracy rate achieved by board-certified echocardiographers. Based on this, Østvik et al. [6] further developed a real-time view classification model, which utilized a dataset containing over 500 patients and 7,000 videos, achieving accuracy rates of 98.3% for single frames and 98.9% for video sequences, respectively. In addition, their system has achieved a real-time processing speed of 4.4 ± 0.3 milliseconds per frame, highlighting its potential in immediate clinical decision-making.

In large-scale applications, Naser et al. [7] trained two-dimensional and three-dimensional CNNS on transthoracic echocardiography (TTE) data from over 900 patients, covering 9 view categories and a

ISSN 2616-5775 Vol. 8, Issue 11: 1-13, DOI: 10.25236/AJCIS.2025.081101

total of 10,269 videos. Their model has an overall accuracy of over 96%, with an area under the curve (AUC) value close to 1.0. At the same time, it also demonstrates strong versatility on the point-of-care ultrasound (POCUS) dataset. To enhance the interpretability of the model, Huang et al. [8] introduced an autoencoder structure inspired by capsule networks into their classification model, making feature mapping deconvolution a decoder for clinical interpretation. This model achieved an average classification accuracy rate of 98.2% and provided enhanced interpretability for clinicians.

For a specific clinical context, Kusunose et al. [9] developed a dataset containing five standard echocardiogram views and trained a CNN classifier, achieving an accuracy rate of 98.1%. This study verified its feasibility in clinical prediction tasks. Wu et al. [10] focused on congenital heart disease (CHD) in children and designed a knowledge extraction framework trained on over 360,000 echocardiographic images. Their system is capable of automatically identifying 23 clinically relevant diagnostic views, with the F1 scores of the majority of these views exceeding 0.90. Recently, Gungor et al. [11] extended view classification by integrating object detection technology, thereby achieving automatic localization and recognition of heart valves, which demonstrated the potential of view classification in supporting more complex clinical applications.

In conclusion, deep learning methods have achieved performance close to that of experts or even surpassing that of humans in the classification of echocardiographic views. However, the high dependence of these methods on labeled datasets and their limited generalization ability under different imaging conditions remain the main bottlenecks. This has prompted researchers to explore methods of self-supervised learning and cross-view representation.

2.2 A comparative study on the classification of echocardiographic views

In recent years, the research on the classification of echocardiographic viewpoints has mainly focused on the following three aspects, as detailed in Table 1:

(1) Significant breakthroughs have been achieved in standard view recognition by using CNNS.

(2) Explore real-time typing and its clinical feasibility;

(3) Integrate interpretability and knowledge distillation methods to enhance its practicality and scalability.

Table 1. Comparison of existing classification studies of echocardiographic views

Work (Authors, Year)	**Dataset size**	**Method**	**Result**
Madani et al., 2018[5]	267 pieces of TTE 15 standard views (12 videos +3 static images)	CNN classification model	The classification accuracy rate of video views is 97.8%. The accuracy rate of static images is 91.7%. The expert accuracy rate is 70% to 84%
Østvik et al., 2019[6]	Over 500 patients; 7000 videos ;Seven views.	CNN, real-time inference	The single-frame accuracy rate is 98.3% ± 0.6%. The sequence accuracy rate is 98.9% ± 0.6%. Real-time performance of 4.4 ms per frame
Kusunose et al., 2020[9]	340 patients; Five standard views; 17,000 images	CNN+ Cross-validation	The test accuracy rate is 98.1%. The error rate is 1.9%, which has no impact on EF prediction
Huang et al., 2022 [8]	26,465 images, 29 types of views	CNN+ Capsule Network Decoder	The average classification accuracy rate is 98.2%. The visualization of interpretability has been verified by experts

ISSN 2616-5775 Vol. 8, Issue 11: 1-13, DOI: 10.25236/AJCIS.2025.081101

Wu et al., 2022 [10]	Patient 3772 367,571 images; 23 children's CHD views	CNN+ Knowledge Distillation	Most views have F1≥0.90; Support automatic diagnosis of CHD in children
Gungor et al., 2023 [11]	A cardiac ultrasound dataset containing 10 standard views	CNN classification + valve localization	The accuracy rate of view classification is relatively high. The detection and positioning of valve bounding boxes were achieved for the first time
Naser et al., 2024 [7]	Patient 909 (10,269 videos, 9 types of views); 229 patient validation set	2D & 3D CNN classification	2D CNN: Accuracy rate 96.8%, AUC=0.997; 3D CNN: Accuracy rate 96.3%, AUC=0.998; The POCUS set has good generalization

3. Method

3.1 Overview of the overall architecture

To enhance the structural understanding ability of the model in the task of echocardiogram view classification, this paper introduces a structure-aware strategy based on the SimSiam framework and, through adaptive modification, proposes a self-supervised representation learning method suitable for echocardiograms. The overall process is shown in Figure 1.

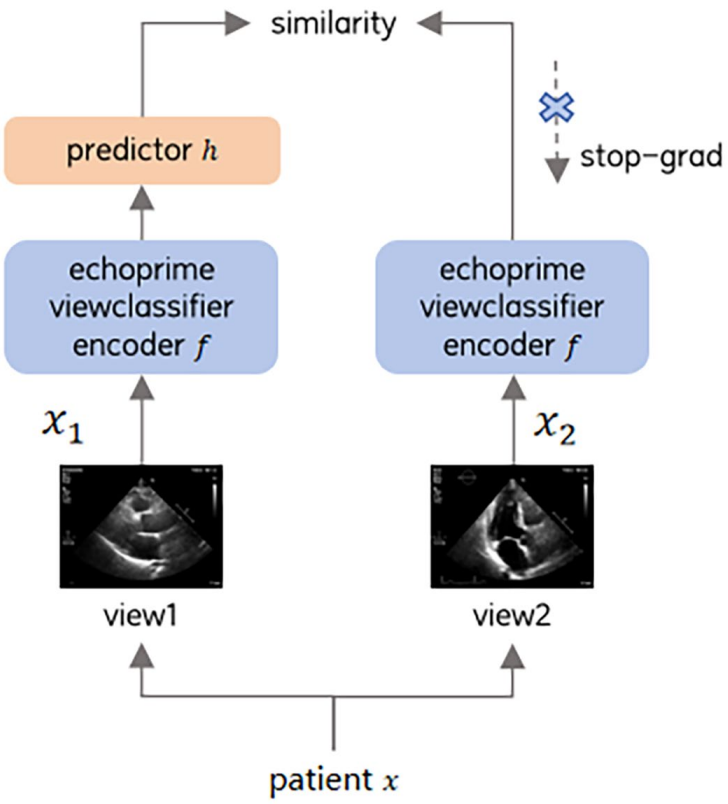

Figure 1. The structure-aware self-supervised learning architecture proposed in this paper

During the training phase, two images of the same patient from different perspectives (denoted as x_1 and x_2) are sampled as input, and features are extracted through an encoder branch with shared

ISSN 2616-5775 Vol. 8, Issue 11: 1-13, DOI: 10.25236/AJCIS.2025.081101

parameters. The left branch is connected to the predictor to achieve feature matching, while the right branch stops the gradient operation to prevent gradient flow, thereby avoiding representation collapse. Ultimately, the model models the consistency of the cross-view structure by maximizing the cosine similarity between the prediction vector and the target vector.

Unlike the traditional SimSiam method that relies on random reinforcement to construct positive sample pairs, the "same patient view pairs" designed in this paper have stronger medical semantics. This design can guide the model to capture the consistency of the heart under different cross-sections, thereby enhancing the discriminability and robustness of the representation.

3.2 Input construction strategy: same patient view pairs

In the framework of self-supervised learning, the strategy for constructing positive sample pairs has a crucial impact on the representational ability of the final model. In the original design of SimSiam, positive sample pairs were constructed by using two enhanced versions of the same image (such as cropping, flipping, or color distortion) to encourage the model to learn the invariance of image perturbations. However, this enhancement method has two significant limitations in echocardiography tasks:

(1) Medical semantic weakening: Conventional enhancement operations may destroy the anatomical structure information in the image, thereby causing the features learned by the model to deviate from the true structural semantics.

(2) Insufficient view diversity: Enhanced views typically exhibit "local variations" and are unable to effectively simulate the real anatomical changes of the heart under different cross-sections.

In response to the above problems, this paper designs a positive sample construction strategy based on pairs of images from different view of the same patient. Specifically, when constructing each training sample, two images from different perspectives (such as PLAX, A2C, A3C, etc.) of the same patient's echocardiography were selected as positive sample pairs(x_1, x_2). Although these two images have significant differences in spatial distribution and visible areas, they are essentially different observations of the same heart structure and have semantic consistency.

Through this strategy, the model is guided to learn the structural commonalities among different echocardiographic views during the self-supervised training process, thereby enhancing its multi-angle understanding ability of cardiac anatomy. Furthermore, this method is more in line with clinical practice because doctors usually comprehensively assess a patient's heart condition from multiple perspectives.

At the implementation level, to ensure the consistency of the samples and the reliability of the labels, we screened out patient samples with multiple high-quality view images from the original dataset. This move aims to ensure the basis of structural consistency between each pair of images and strictly group them according to patient ids to prevent semantic deviations caused by cross-patient sampling.

In conclusion, compared with traditional data augmentation methods, the same-patient view construction strategy proposed in this paper not only effectively retains the semantics of medical structure but also guides the model to model the invariance of cardiac structure across multiple views at a deeper level. This provides more powerful feature support for subsequent downstream classification tasks.

3.3 Model architecture adaptation and modification

To better adapt to the task of echocardiographic view classification, based on the analysis of the existing self-supervised framework, this paper designs an improved dual-branch architecture. On the one hand, this design draws on the "project-prediction" mechanism commonly used in contrastive learning; On the other hand, it combines optimizations for specific tasks of echocardiographic images, thereby forming a structure-aware self-supervised architecture that better meets the needs of medical scenarios.

3.3.1 Encoder design: ConvNeXt Basic backbone

In this study, we selected the ConvNeXt Base network as the backbone of the encoder for feature extraction. Compared with the traditional ResNet series, ConvNeXt adopts a wider convolution kernel

ISSN 2616-5775 Vol. 8, Issue 11: 1-13, DOI: 10.25236/AJCIS.2025.081101

and a deeper normalization mechanism, which enables it to effectively simulate the blurred edges and global structural information in echocardiographic images, thereby better adapting to the low contrast characteristics of medical images. To make full use of the existing pre-training parameters, we transiently initialized ConvNeXt and redesigned the subsequent modules.

3.3.2 Redesign the projection and prediction modules

After obtaining the feature representation from the encoder, this paper introduces two types of multi-layer perceptrons (MLPS) :

Projection Head: It adopts a three-layer structure, with a hidden layer dimension of 1024 and an output dimension of 2048. The last layer removes the ReLU activation to ensure that the output vector is suitable for similarity calculation.

Predictor: It adopts a three-layer bottleneck structure, with both input and output dimensions of 2048, while the hidden dimension is set at 512.By introducing batch normalization and nonlinear activation functions, the predictive ability has been enhanced and the occurrence of degradation has been effectively prevented.

Unlike general self-supervised frameworks, this paper conducts in-depth optimization of the structure of the predictor. This is because cross-view matching is more challenging than matching between enhanced versions of the same image, and thus requires stronger feature alignment and prediction capabilities.

3.3.3 Stop-gradient mechanism and Training Process Adjustment

In order to prevent feature collapse in the two-branch network during training, this paper introduces the stop gradient operation in the right branch. This design ensures that the feature vector of this branch does not participate in the backpropagation of the gradient, but only serves as a comparison target. Although similar ideas exist in other frameworks, this design is particularly important in our research scenario, where cross-view inputs exhibit greater variability and thus are more likely to lead to training instability.

In summary, the proposed model architecture does not rely on direct reuse of a single existing framework. Instead, it redesigns and modifies the structure based on the analysis of effective mechanisms. By selecting a ConvNeXt backbone that is more suitable for echocardiographic image processing, deepening the predictor structure, and integrating the stopping gradient mechanism, we construct a structure-aware self-supervised network that can effectively model cross-view consistency and provide a solid feature representation foundation for subsequent classification tasks.

3.4 Loss Function and Training Objective

To effectively drive the model to learn discriminative view-invariant features under unsupervised conditions, this paper adopts the symmetric similarity maximization strategy proposed by SimSiam as the training objective. This method does not require negative sample support. By introducing the prediction module and the gradient stop mechanism, this model can achieve the alignment of the underlying structural semantics among different views.

Let the input image pairs$(\mathrm{x}_1, \mathrm{x}_2)$, represent two different views of the same patient. After processing by the encoder f, projector g and predictor h with shared weights, the following vector representation can be obtained:

$$z_1 = g\big(f(x_1)\big), p_1 = h(z_1) \tag{1}$$

$$z_2 = g\big(f(x_2)\big), p_2 = h(z_2) \tag{2}$$

During the training process, gradient backpropagation only updates the predictor of one branch, while stopping the gradient operation blocks the projection vector of the other branch. Specifically, for$p_1$and z_2, our goal is to maximize the cosine similarity between them:

$$D(p_1, z_2) = -\frac{p_1}{\|p_1\|_2} \cdot \frac{z_2}{\|z_2\|_2} \tag{3}$$

To ensure the symmetry and stability of the training process, we maximize the similarity of the predictions in both directions and take the average as the final loss function:

ISSN 2616-5775 Vol. 8, Issue 11: 1-13, DOI: 10.25236/AJCIS.2025.081101

$$\mathcal{L}_{total} = \frac{1}{2}D(p_1, z_2) + \frac{1}{2}D(p_2, z_1) \tag{4}$$

This method is defined for each image, and the total loss is averaged over all images. Its minimum possible value is -1.

One of the important components of this method is to stop the gradient operation (see Figure 1). We achieve this by modifying equation (3) :

$$D\big(p_1, stopgrad(z_2)\big) \tag{5}$$

This means that in this term, z_2 is regarded as a constant. Similarly, Equation (4) is also implemented in a similar form:

$$\mathcal{L}_{total} = \frac{1}{2}D\left(p_1, stopgrad(z_2)\right) + \frac{1}{2}D\left(p_2, stopgrad(z_1)\right) \tag{6}$$

This design has the following advantages: Firstly, it eliminates the need for negative samples and avoids the construction of large negative sample sets commonly used in contrastive learning, thereby improving the training efficiency. Secondly, by leveraging the structural consistency among different views from the same patient, semantic alignment between views is forcibly achieved, guiding the model to learn more robust features under view changes. Finally, driven by structural consistency, the loss term implicitly encourages the model to extract consistent structural information in different views, thereby enhancing the anatomical representation of the heart.

To enhance the numerical stability during the training process, l_2 normalization was applied to the prediction vector and the projection vector before calculating the cosine similarity. In addition, to prevent the predictor from falling into regression resolution (such as identity mapping or constant output), a nonlinear activation function and dropout (p=0.2) were introduced in the predictor module to enhance the model's sensitivity to differences in image structure.

Based on approach[21], when $\mathrm{p_i} = -\mathrm{z_j}$, loss function $\mathrm{D}(\mathrm{p_i}, \mathrm{z_j}) = -(-1)$, the largest loss of 1; When $\mathrm{p_i} = \mathrm{z_j}$ is a perfect match, the loss reaches its minimum value of -1. Therefore, the range of values for this loss function is [-1, 1]. The optimization objective is to minimize $\mathcal{L}_{\mathrm{total}}$, that is, to maximize structural consistency among cross-view vectors.

This training objective, while maintaining the efficiency and simplicity of SimSiam, further introduces a similarity supervision mechanism driven by structural prior. This mechanism provides a stable optimization objective for extracting discriminative structural features and supporting downstream view classification tasks.

3.5 Implementation Details and Training Procedure

3.5.1 Data preprocessing and input construction

In each training epoch, the model randomly selects two echocardiographic images from different perspectives of the same patient as the front pair $(\mathrm{x_1}, \mathrm{x_2})$ for training the structural alignment representation. This input structure reflects the assumption of structural invariance in echocardiographic images, that is, despite differences in imaging angles, the anatomical structure of the heart remains consistent. Therefore, the model should be capable of learning such cross-perspective semantic consistency.

To fully tap into the data potential of each patient, we have built a customized dataset - EchoViewDataset. This dataset supports multiple samplings of different view combinations for each patient, thereby significantly expanding the amount of training data. The specific data loading logic is as follows:

(1) Data organization is based on the patient directory, and each patient folder contains multiple subfolders with different views (for example, PLAX, A4C, etc.);

(2) For each training sample, two different views are randomly selected from the same patient, and an image is randomly chosen from each view to form a direct alignment.

(3) To ensure data quality, patients with less than two view categories or missing images are automatically excluded;

ISSN 2616-5775 Vol. 8, Issue 11: 1-13, DOI: 10.25236/AJCIS.2025.081101

(4) The number of pairs of samples for each patient is controlled by the parameter samples_per_patient (default is 40), which strikes a balance between training scale and diversity.

Through the above methods, we can effectively enhance the richness and representativeness of the data required for model training.

For image enhancement, to avoid interference with the anatomical structure of the heart, we have designed a lightweight enhancement strategy, which includes geometric transformation, slight image perturbation, normalization processing and tensor transformation.

Finally, the enhanced image pairs are input into the shared encoder in the form of$(x_1, x_2)$for similarity learning. This input construction strategy combines the advantages of medical prior knowledge and self-supervised learning, not only improving the stability of the training process but also enhancing the model's perception ability of structural invariance.

3.5.2 Training parameter Settings

We use the SGD optimizer for pre-training. The learning rate is set according to the linear scaling rule lr×BatchSize / 256 (see linear scaling [22]), with an initial learning rate of 0.05, and cosine decay scheduling [23,15] is used. The weight decay is set to 0.0001, and the momentum value of SGD is 0.9. In this experiment, we used the NVIDIA A100-PCIE-40GB GPU, set the batch size to 64, and trained the model for 100 epochs.

4. Results

4.1 Experimental Setup

Experimental environment: NVIDIA A100 40GB GPU, Python 3.9.12, PyTorch 2.1.2, CUDA 12.1 / CuDNN 8.9.2.

Dataset segmentation: In this study, the dataset is divided into a training set and a validation set, with a specific segmentation ratio of 7:3. During this process, all patient samples from the validation set were excluded from the training set to ensure patient-level independence.

Training parameters: This study employs a stochastic gradient descent (SGD) optimizer, with the batch size set at 64. The training process lasts for 100 cycles. The initial learning rate is set at 0.05, and the cosine attenuation scheduling strategy is used. The loss function selects the maximization of symmetric cosine similarity as its objective. All models are implemented in the backbone network with ConvNeXt Base as the encoder.

4.2 Experimental Comparison Design

To systematically evaluate the effectiveness of the proposed structure-aware self-supervised pre-training method in the classification of echocardiographic views, we designed three sets of comparative experiments, aiming to demonstrate the performance evolution from traditional supervised learning methods to our self-supervised learning strategy. All experiments were conducted on the same dataset and downstream classification configuration to ensure the comparability of the results. The overall experimental plan is shown in Figure 2, and the specific description is as follows.

(1) Experiment 1: View Classification Model Based on ConvNeXt

In this experiment, the ConvNeXt Base model pre-trained with large-scale labeled data was used for the classification of echocardiographic views. This encoder has been specially trained for view classification and thus possesses a strong discriminative ability. Considering the robustness demonstrated by the encoder in visual perception, we only fine-tuned the classification head while keeping the encoder parameters unchanged. This setting simulates common transfer learning strategies in real-world scenarios and serves as an important benchmark reference for supervising the performance evaluation of transfer learning.

(2) Experiment 2: ConvNeXt Base+Structure-aware Self-supervised Pre-training

In this experiment, we adopt ConvNeXt Base as the encoder and conduct further pre-training in combination with the proposed structure-aware self-supervised method. By introducing structural matching and consistency enhancement tasks, the aim is to guide the encoder to learn feature representations with semantic structure-aware characteristics. After the self-supervision phase ended,

ISSN 2616-5775 Vol. 8, Issue 11: 1-13, DOI: 10.25236/AJCIS.2025.081101

we made a comprehensive fine-tuning of the entire network (including the encoder and classification heads) to adapt to the downstream view classification tasks. This experiment aims to evaluate whether the proposed structure-aware pre-training can further enhance performance on the basis of the existing pre-training.

(3) Experiment 3: Random initialization of ConvNeXt Base + self-supervised pre-training

To further verify the independence and generalization ability of the proposed self-supervised method, in this experiment, the ConvNeXt Base encoder was randomly initialized, and structure-aware self-supervised pre-training was implemented from scratch. Subsequently, comprehensive fine-tuning was carried out for the downstream classification tasks. This experiment eliminates the possible influence of supervised pre-training, thereby providing a clearer basis for evaluating the intrinsic effectiveness of the proposed structure-aware self-supervised method.

To ensure the fairness and scientific rigor of the comparison experiments, we adopted differentiated fine-tuning strategies based on the pre-training objectives and feature semantics of each model:

(1) Experiment 1 (ConvNeXt Base View Classification Model): Since the encoder has been pre-trained on the supervised task and has fully learned the classification discriminant features, only the classification head is fine-tuned after the transfer, while keeping the encoder parameters unchanged.

(2) Experiments 2 and 3 (Structure-aware Self-supervised Model) : The encoders pre-trained using the structure-aware self-supervised strategy mainly learn structural representations rather than direct view discriminative features. Therefore, in downstream tasks, the entire network was comprehensively fine-tuned to achieve an effective mapping from structural representation to semantic categories.

This strategic design not only maximizes the performance potential of each model under its own conditions, but also ensures the comparability when evaluating transferability and feature representation capabilities.

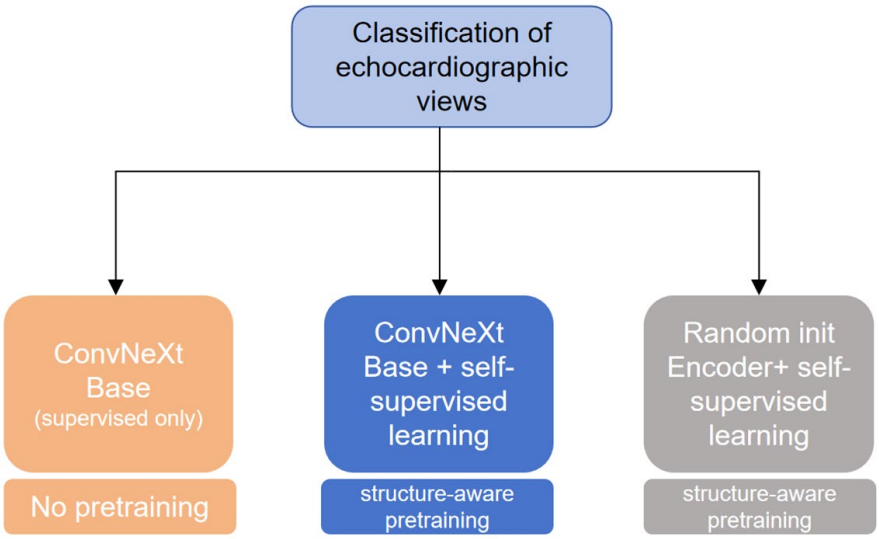

Figure 2. The process structure diagrams of three groups of comparative experiments

This figure 2 shows the process structure of three groups of experiments, covering model sources, pre-training methods, fine-tuning strategies, and their adaptation processes in downstream classification tasks. Among them, the orange path represents the ConvNeXt Base supervised model where only the classification head is fine-tuned directly; The blue path represents a ConvNeXt with additional structure-aware self-supervised pre-training, which is then fully fine-tuned. The grey path describes a randomly initialized ConvNeXt model, which is trained using the same self-supervised strategy and then fully fine-tuned.

4.3 Experimental Results and Performance Analysis

This section presents the classification performance evaluation of three comparative experiments conducted on the validation set. We randomly selected 100 echocardiograms from the validation set to evaluate their accuracy, confusion matrix, ROC curve and AUC indicators. The results are presented in

ISSN 2616-5775 Vol. 8, Issue 11: 1-13, DOI: 10.25236/AJCIS.2025.081101

both tabular and visualized forms, supplemented by quantitative and qualitative analyses to comprehensively evaluate the effectiveness of the proposed method.

4.3.1 Comparison of Classification Accuracy

Table 2 summarizes the classification accuracy (Val Acc) of the three experiments on the validation set. The main findings are as follows:

(1) ConvNeXt Base (Supervised pre-training, fine-tuning the classification header only) :The accuracy rate of this model has reached 89.75%. Due to the encoder's features being biased towards supervised classification tasks and lacking adaptability to various structural changes, its performance is limited, resulting in a relatively low accuracy rate.

(2) ConvNeXt Base + Structure-aware Self-supervised Pre-Training: On the validation set, this model achieved an accuracy rate of 98.25%, significantly outperforming classification models that directly adopt supervised learning. This result indicates that the proposed structure-aware self-supervised strategy effectively enhances the modeling of structural consistency among different echocardiographic views, thereby improving the robustness during the classification process.

(3) Random initialization + structure-aware self-supervised pre-training: This method achieved an accuracy rate of 96.50%, indicating that even without supervised pre-training, this method can still obtain powerful feature representations through structure-aware self-supervised learning. This highlights its excellent generalization ability.

In conclusion, the proposed self-supervised method not only fully exploits the advantages of the pre-trained ConvNeXt encoder, but also can learn discriminative structural representations from scratch during random initialization, thereby significantly improving the accuracy of echocardiographic image classification.

Table 2. Comparison of validation accuracy among the three experiments (%)

Experiment ID	Model Configuration	Backbone Network	Pre-training method	Fine-tuning strategy	Verification Accuracy (Val Acc)
Exp. 1	ConvNeXt Base (supervised classification)	ConvNeXt Base	None	Classification head only	89.75
Exp. 2	ConvNeXt Base + SSL	ConvNeXt Base	Structure-aware	Full fine-tuning	98.25
Exp. 3	Random Init + SSL	ConvNeXt Base	Structure-aware	Full fine-tuning	96.50

4.3.2 Confusion Matrix Analysis

Figure 3 shows the confusion matrices of three groups of experiments. The research results are summarized as follows:

(1) Experiment 1 shows that there is a significant confusion between certain categories (such as A2C and PLAX), which reflects that the supervised classification model has certain problems in cross-perspective discrimination.

(2) The results of Experiment 3 show that the degree of confusion is significantly reduced compared with the supervised baseline. However, misclassification still occurs between some anatomically similar perspectives. Overall, although there is still slight confusion in specific categories, the performance of Experiment 3 is better than that of Experiment 1.

(3) Experiment 2 achieved the best performance in all categories and significantly reduced the confusion rate. This result indicates that the proposed structure-aware self-supervised strategy effectively learns the invariance of cross-view structures, thereby improving the classification accuracy.

ISSN 2616-5775 Vol. 8, Issue 11: 1-13, DOI: 10.25236/AJCIS.2025.081101

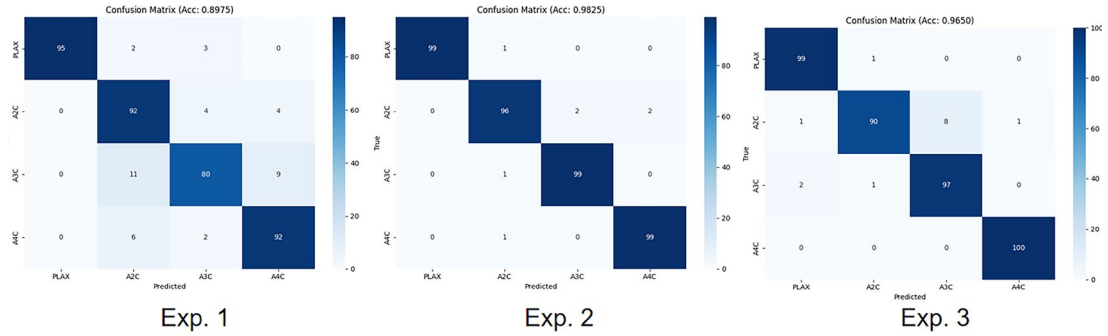

Figure 3. Confusion matrices of the three experiments

4.3.3 ROC Curve and AUC Analysis

Figure 4 shows multiple types of ROC curves and their corresponding AUC indicators under three experimental setups:

(1) Experiment 1: The overall ROC curve deviates significantly from the ideal situation. The AUC values of some view categories are only close to 0.90, indicating that their classification ability is limited.

(2) Experiment 2: The ROC curve was almost perfectly aligned in the upper left corner, and the AUC values of each category exceeded 0.98, demonstrating excellent discriminative performance. This further verified the effectiveness and robustness of the method proposed in this paper in the classification of echocardiographic views.

(3) Experiment 3: The ROC curve as a whole was closer to the ideal upper left boundary, and the AUC values of most categories exceeded 0.95, highlighting the advantages brought by the structure-aware self-supervised method.

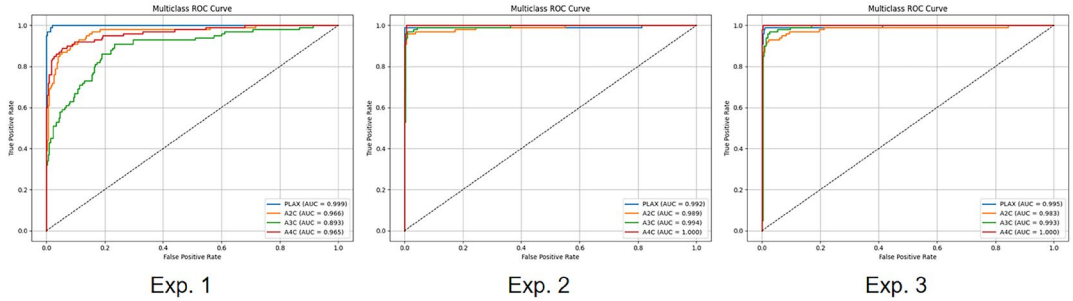

Figure 4. Multi-class ROC curves and AUC comparison of the three experiments

Based on the results of Table 2, Figure 3 and Figure 4, the following conclusions can be drawn:

(1) Supervised ConvNeXt Base classifiers have obvious limitations in terms of overall accuracy and category-level discriminability, and thus are not very suitable for complex clinical scenarios.

(2) Random initialization combined with self-supervised pre-trained models can effectively learn feature representations under unsupervised conditions, and their classification performance is significantly better than that of pure supervised baselines.

(3) The combination of ConvNeXt Base and self-supervised pre-trained models consistently shows the best results in terms of accuracy, confusion matrix separability, and ROC-AUC metrics. This discovery confirms that the proposed structure-aware self-supervised method effectively utilizes the prior of structural consistency, thereby enhancing the robustness and generalization ability across views.

5. Conclusion

5.1 Research Conclusions

Aiming at the basic clinical task of classifying echocardiographic views, this paper proposes an

ISSN 2616-5775 Vol. 8, Issue 11: 1-13, DOI: 10.25236/AJCIS.2025.081101

improved method based on structure-aware self-supervised learning. Starting from the clinical practical challenges such as the diversity of echocardiographic views, the ambiguity of anatomical structures, and the high cost of manual annotation, we have carried out systematic exploration and improvement. The main conclusions are as follows:

(1) A construction strategy for different views of the same patient was proposed.By using different views from the same patient as positives instead of the positives in traditional augment-based self-supervised learning, we effectively introduce medical structure priors. This strategy enables the model to learn structural consistency across views, thereby better adapting to echocardiographic diagnosis in the real world and providing a strong representational foundation for downstream tasks.

(2) A structure-aware self-supervised learning framework was designed.Inspired by the SimSiam mechanism, we integrated the ConvNeXt basic backbone and constructed a customized self-supervised representation learning system for application in echocardiographic images. After combining the stop gradient mechanism and symmetric loss design, representation collapse was effectively prevented, ensuring the stability of the feature learning process.

(3) Significantly improved the performance of view classification.The experimental results show that, compared with directly using the supervised ConvNeXt Base model, this method improves the verification accuracy by nearly 10 percentage points. Compared with the randomly initialized self-supervised model, this method maintains an advantage of more than 1.5 percentage points. Further analysis of the confusion matrix and ROC-AUC confirmed that our method significantly enhanced class discriminability and overall robustness, especially in differentiating highly similar views (such as A2C and A3C).

(4) It has strong potential for clinical application.This study, by introducing structure-aware self-supervised learning, not only enhances the classification performance but also provides a new approach to alleviate the problem of limited labeling. The proposed framework can effectively utilize large-scale unlabeled echocardiographic data with limited labels, providing a more efficient and reliable human-intelligence auxiliary tool for clinical decision-making.

In conclusion, the contrastive Cross-View representation learning method proposed in this paper has achieved remarkable results in the classification of echocardiographic images. This method not only outperforms existing technologies in terms of accuracy and robustness, but also shows a high degree of consistency with the features of medical imaging. This study lays a solid foundation for more complex intelligent diagnostic tasks in future echocardiography.

5.2 Future Work

The focus of future research will be on enhancing the generalization ability of echocardiographic image classification. On the one hand, cross-center and cross-device validation methods should be actively explored to enhance the applicability of this technology in different clinical Settings. On the other hand, combining multimodal information such as echocardiographic video sequences and clinical examination reports is conducive to further deepening the understanding of cardiac structure and function. Overall, on the basis of maintaining a high classification accuracy, improving the robustness and interpretability of the model will become an important direction for future research.

Acknowledgements

This work was supported by the One Health Interdisciplinary Research Project of Ningbo University (no. HY202409) and the S&T Program of Hebei No. 22377774D.

References

[1] Benjamin EJ, Muntner P, Alonso A, et al. Heart Disease and Stroke Statistics—2019 Update: A Report From the American Heart Association. 2019;139(10):e56-e528.
[2] Rudski LG, Lai WW, Afilalo J, et al. Guidelines for the Echocardiographic Assessment of the Right Heart in Adults: A Report from the American Society of Echocardiography: Endorsed by the European Association of Echocardiography, a registered branch of the European Society of Cardiology, and the Canadian Society of Echocardiography. Journal of the American Society of Echocardiography. 2010;23(7):685-713.

ISSN 2616-5775 Vol. 8, Issue 11: 1-13, DOI: 10.25236/AJCIS.2025.081101

[3] Nagueh SF, Smiseth OA, Appleton CP, et al. Recommendations for the Evaluation of Left Ventricular Diastolic Function by Echocardiography: An Update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging. European Heart Journal - Cardiovascular Imaging. 2016;17(12):1321-1360.

[4] Lang RM, Badano LP, Mor-Avi V, et al. Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging. 2015;16(3):233-271.

[5] Madani A, Arnaout R, Mofrad M, et al. Fast and accurate view classification of echocardiograms using deep learning. npj Digital Medicine. 2018/03/21, 2018;1(1):6.

[6] Østvik A, Smistad E, Aase SA, Haugen BO, Lovstakken L. Real-Time Standard View Classification in Transthoracic Echocardiography Using Convolutional Neural Networks. Ultrasound in Medicine & Biology. 2019/02/01/, 2019;45(2):374-384.

[7] Naser JA, Lee E, Pislaru SV, et al. Artificial intelligence-based classification of echocardiographic views. European Heart Journal - Digital Health. 2024;5(3):260-269.

[8] Huang M, Lin W, Chen Y, et al. Explainable deep neural network for echocardiography view classification. European Heart Journal - Cardiovascular Imaging. 2022;23(Supplement_1)

[9] Kusunose K, Haga A, Inoue M, Fukuda D, Yamada H, Sata M. Clinically Feasible and Accurate View Classification of Echocardiographic Images Using Deep Learning. Biomolecules. 2020; 10(5):665.

[10] Wu L, Dong B, Liu X, et al. Standard echocardiographic view recognition in diagnosis of congenital heart defects in children using deep learning based on knowledge distillation.Frontiers in Pediatrics. 2022;9:770182.

[11] Gungor DG, Rao B, Wolverton C, et al. View classification and object detection in cardiac ultrasound to localize valves via deep learning.arXiv preprint arXiv:2311.00068 (2023).

[12] Smistad E, Østvik A, Salte IM, et al. Real-time automatic ejection fraction and foreshortening detection using deep learning.IEEE transactions on ultrasonics, ferroelectrics, and frequency control. 2020;67(12):2595-2604.

[13] Tian Y, Xie L, Zhang X, et al. Semantic-aware generation for self-supervised visual representation learning.arXiv preprint arXiv:2111.13163 (2021).

[14] Le-Khac PH, Healy G, Smeaton AFJIA. Contrastive representation learning: A framework and review. IEEE Access. 2020;8:193907-193934.

[15] Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. PmLR; 2020:1597-1607.

[16] He K, Fan H, Wu Y, et al. Momentum contrast for unsupervised visual representation learning.Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020:9729-9738.

[17] Grill J-B, Strub F, Altché F, et al. Bootstrap your own latent-a new approach to self-supervised learning.Advances in neural information processing systems. 2020;33:21271-21284.

[18] Chen X, He K. Exploring simple siamese representation learning.Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021:15750-15758.

[19] Zhou Z, Sodha V, Rahman Siddiquee MM, et al. Models genesis: Generic autodidactic models for 3d medical image analysis. Springer; 2019:384-393.

[20] Chen S, Ma K, Zheng YJapa. Med3d: Transfer learning for 3d medical image analysis. arXiv preprint arXiv:1904.00625 (2019).

[21] Grill J-B, Strub F, Altché F, et al. Bootstrap your own latent-a new approach to self-supervised learning.Advances in neural information processing systems. 2020;33:21271-21284.

[22] Goyal P, Dollár P, Girshick R, et al. Accurate, large minibatch sgd: Training imagenet in 1 hour. arXiv preprint arXiv:1706.02677 (2017).

[23] Loshchilov I, Hutter FJapa. Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016).