QueryCLR: a Query-based Contrastive Visual Representation Learning for Echocardiogram Classification

Zhen Li[1], Hao Wang[2, *], Yingxiu Chang[3], Weijing Qiu[1], Yongqiang Cheng[4]

[1] Jiangsu Reckon Medical Intelligence Company Limited, Jiangsu Province, China
[2] Department of Echocardiography, State Key Laboratory of Cardiovascular Disease, Fuwai Hospital, National Center for Cardiovascular Diseases, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China
[3] School of Computer Science, University of Hull, Cottingham Road, Hull, HU6 7RX, UK
[4] Faculty of Technology, University of Sunderland, St. Peter's Campus, Sunderland, SR6 0DD, UK

***Abstract*—Echocardiography is a leading cardiac imaging technique that requires strong expertise when performing an analysis, and manual labeling is challenging due to various acquisition devices. This research uses an innovative query-based contrast learning visual representation methodology called QueryCLR. By integrating CNN's precise feature extraction with transformer models' broad contextual understanding, QueryCLR substantially enhances the stability and accuracy of image analysis. Our experimental analysis, conducted on 12 different two-dimensional echocardiogram datasets, validates QueryCLR's effectiveness. The method achieved a maximum accuracy of 79.4%, showing a 1.2% superiority over existing self-supervised models. Moreover, by just using 10% of labeled data via transfer learning, QueryCLR achieved 84.4% accuracy, surpassing the supervised learning model by 0.9%. These results highlight the promising potential of self-supervised learning in enhancing accuracy and efficiency in cardiac disease identification through echocardiographic classification while providing new insights for future research.**

Keywords—Self-supervised learning, Contrastive Learning, Vision Transformer, Visual Representation, Echocardiogram classification

I. Introduction

Echocardiography is a widely used heart imaging technique that captures dynamic images of the heart's structure and function through high-frequency sound waves. This non-invasive procedure helps observe and evaluate the structure and function of the heart, making it an important tool for diagnosing and monitoring heart disease. Echocardiography can provide detailed of the two-dimensional heart and detect early signs of heart disease, such as heart valve disease and myocardial infarction. However, given the intricate nature and technicality of echocardiograms, a high level of expertise is required by physicians for accurate analysis, while acknowledging that variations may exist in the images obtained from different devices. With the rapid development of Artificial Intelligence (AI) technology, AI has become a powerful tool in the field of medical image analysis. Deep learning as an important branch of AI, has made remarkable progress in medical image recognition and classification[1],[2]. In echocardiography, deep learning can automatically identify and analyze heart structures, abnormalities, and lesions to provide faster, more accurate, and consistent diagnostic results. This has great potential to improve the quality of medical care for patients and reduce the time spent waiting for a diagnosis.

Currently, echocardiogram research primarily relies on supervised learning methods. However, this approach requires extensive manual annotation by professionals with specialized expertise, resulting in significant labor and time. Given the complex nature of medical domains, the limited labels in supervised learning prove a challenge to be collected.

Self-supervised learning enables models to learn useful feature representations by automatically generating labels from input features. Self-supervised learning obviates the need for large-scale labeled datasets, significantly reducing the demand for labeled data. This approach is highly effective when labeled data is scarce or costly. For these reasons, self-supervised learning methods are becoming increasingly important in addressing the challenges of limited labeling data [3]. By leveraging unlabeled data, self-supervised learning enhances the model's generalization performance, allowing it to more easily adapt to different domains, distributions, or perspectives within the data. A key stream of self-supervised learning is contrastive learning. Contrastive learning is a self-supervised learning method that aims to learn valuable feature representations in the data by training the model to differentiate between various versions of the same sample, for example, positive versus negative samples[4]. Unlike most self-supervised instance-based methods that limit the diversity of negative instances by categorizing or clustering data instances, contrastive learning focuses on comparing different instances to understand their similarities and differences[3],[5]. This makes self-supervised contrastive learning a powerful tool in modern machine learning, particularly in fields where annotated data is scarce or expensive to obtain [6].

Convolutional Neural Network (CNN)[7] is a traditional model in computer vision, particularly adaptive at perceiving local information within images, which gives them an advantage in tasks like image classification and object recognition. However, due to their inherent spatial induction bias and reliance on prior knowledge, CNN may face limitations in tasks requiring a broader context understanding or scenarios where the input data deviates significantly from the training data. The Vision Transformer (ViT)[8] is a model based on a self-attention mechanism. This mechanism emphasizes more attention to global feature information in feature extraction, leading to improved performance in image classification. However, due to its high computational demands and training requirements on large data sets, ViT needs high computational costs and necessitates complex training setups.

By synergistically integrating the robust capabilities of both Convolutional Neural Networks (CNN) and Vision Transformers (ViT), this novel approach effectively addresses the inherent challenge of capturing intricate image features in echocardiography. This fusion enhances

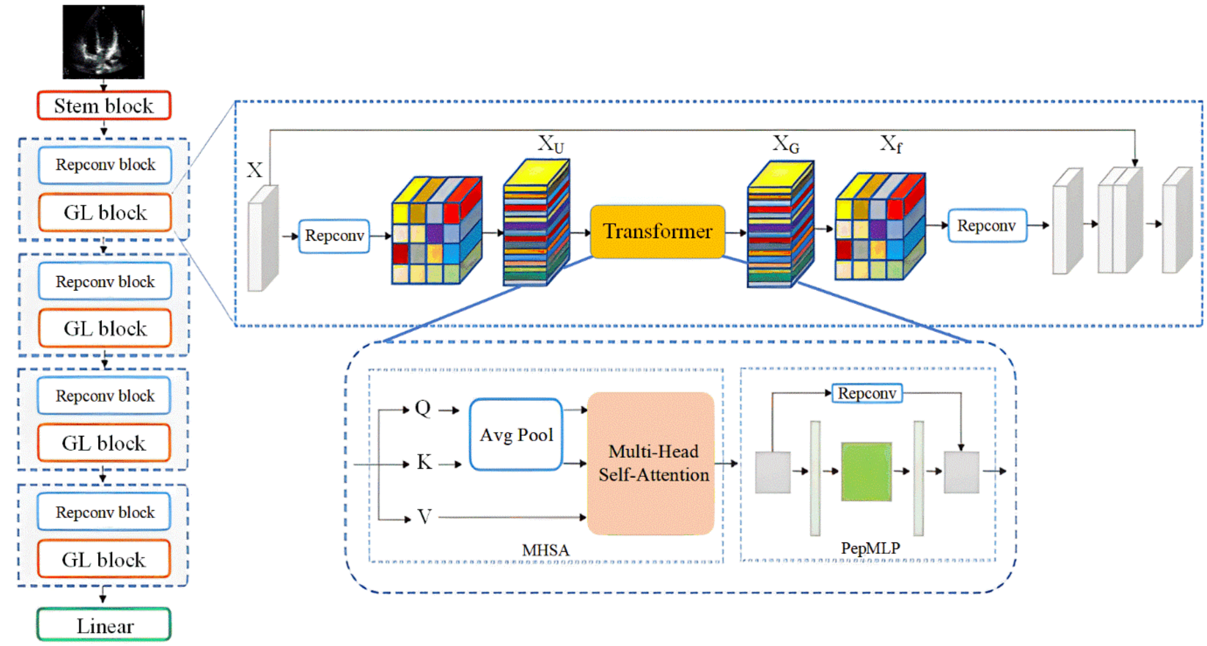


Figure 1. Overview of Global-and-local Network (GLNet)

the model's interpretive prowess, enabling it to accurately decipher the complex and dynamic nature of echocardiographic data, thereby significantly elevating the precision and dependability of classification models in cardiac imaging.

In this paper, a visual representation network called QueryCLR is proposed by combining the advantages of convolution and Transformer and utilizing self-supervised learning. The encoder is pre-trained by contrast learning. Perturbations of various instances are generated by query sets of negative instances to address the issue of insufficient diversity. In addition, we proposed a backbone network called Global-Local Network (GLNet) that combines convolution and Transformer architectures. By embedding the Transformer structure into convolution, we incorporate the desired features of CNN into the ViT architecture while retaining the advantages of the Transformer. The main contributions of this work are as follows:

1) This study presents a new dataset that provides echocardiography scanned across 12 categories, providing a valuable resource for further study and understanding of echocardiography.

2) The QueryCLR framework is proposed, which generates diversity in positive samples through query sets, significantly reducing the dependence on data augmentation and improving feature representations.

3) The GLNet model was designed as the encoder for QueryCLR, combining CNN with local perception and Transformers with global modeling capabilities to achieve global representations with spatial inductive bias.

II. Related Research

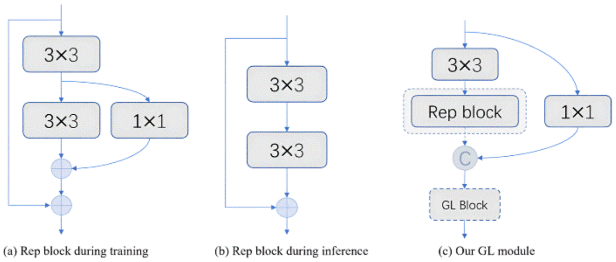
A. Research in Echocardiogram

Recent studies utilizing CNN architectures have made significant advancements in echocardiography. Madani et al[9], Zhang et al[10], and Gao et al[11]. have each employed CNN for various tasks, including view classification and disease diagnosis in echocardiographic imagery. These approaches have demonstrated high accuracy, showcasing the potential of deep learning in medical imaging analysis. However, the application of these algorithms to echocardiography is not without challenges. One major limitation is the reliance on extensive, labeled datasets, which are particularly diverse in echocardiography due to variations in patients, equipment, and scanning parameters[12]. This diversity complicates model training, necessitating algorithms capable of processing a wide range of data inputs[13].

To overcome these hurdles, researchers are investigating self-supervised learning models. These models, less reliant on labeled datasets, have shown promise in other medical imaging areas. For instance, Wilson et al[14]. successfully applied a self-supervised learning algorithm for prostate cancer classification using micro-ultrasound data, and VanBerlo et al[15]. improved model performance through self-supervised pre-training for lung section classification in ultrasound. Despite these advancements, tackling unlabeled data remains intricate due to the complexities of echocardiograms.

B. Self-supervised Learning

Fully supervised network models typically focus on specific tasks. However, in scenarios where specific tasks suffer from a scarcity of data and labels, the effectiveness of such models is significantly limited, as emphasized in previous studies[16, 17]. The SimCLR[3] framework addresses this by applying random transformations to an image, thereby generating two augmented representations. It then seeks to maximize the similarity between these representations to develop a more generalized model. Nonetheless, SimCLR faces limitations due to its uniform dictionary and batch sizes, leading to a lack of diversity for different instances of the same object. Moreover, the challenge of optimizing overly large batches often hinders convergence. MoCo[4] adopts a different approach by utilizing various



(a) Rep block during training (b) Rep block during inference (c) Our GL module

Figure 2. Structure of our GL stage

image enhancement strategies to derive representations. It employs a queue-based dynamic dictionary to accommodate a larger sample size, ensuring a higher quantity of negative samples in each batch. Additionally, it includes a momentum update encoder to uphold key representation consistency amid rapid changes in the encoder, albeit at the cost of slower updates. Dwibedi [18] suggests an alternative method, proposing the use of nearest neighbors to sample from the latent space data. This approach treats these neighbors as positive examples, offering more semantic variation beyond mere data augmentation. SimMM[19] on the other hand, focuses on predicting the original signal by randomly masking certain blocks. These are then encoded and regressed in the masked region using a single-layer prediction head. Collectively, these self-supervised learning methods excel at deriving a generic feature representation that can be seamlessly integrated into a variety of downstream tasks.

Self-attention mechanisms have been widely applied to CNN in vision tasks[20]. Traditional CNN relies on convolution operations that process spatially adjacent input features with fixed fusion weights. This local processing is efficient but inherently limited in capturing long-range dependencies within the input data. Self-attention, on the other hand, allows the network to weigh the importance of different parts of the input data, regardless of their spatial proximity. This mechanism enables the model to focus more on relevant features and less on irrelevant ones, enhancing the network's ability to capture complex patterns and dependencies[21, 22]. Convolutional neural networks use fixed fusion weights for spatially adjacent input features, whereas SENet[23] improves the network's performance by capturing long-range dependencies through global attention[24]. Mobile-Former[25] combines the parallel design of MobileNet[26] and Transformer to extract pixel-level local features using convolution and then encodes global features using Transformer to achieve a bidirectional fusion of local and global features. BoTNet[27] introduces a simple yet powerful backbone that substitutes spatial convolution with global self-attention in the final three bottleneck blocks of ResNet[16], resulting in impressive performance gains. These networks all incorporate geometric priors in adaptive weight fusion and play a significant role in the recognition task.

Self-supervised learning provides an efficient solution to the complexity of echocardiography. By automatically extracting the deep features of heart images, it allows the full utilization of unlabeled data and effectively addresses the challenges of high heart dynamics, large image variation, and complex structure. Reduce the reliance on expert labeling, reduce costs, and improve the ability to generalize in the face of disease diversity and individual differences. Self-supervised learning enhances the accuracy of echocardiogram classification and brings a new and efficient way for the diagnosis of heart disease.

III. Methodology

In this section, we propose a query-based framework for contrastive learning of visual representation (QueryCLR) to find the most similar samples as positive pairs from an augmented collection of echocardiogram cutout datasets to obtain two correlated views. After training, we freeze the projection head and utilize only the encoder GLNet and the representation to classify different echocardiograms.

A. Global-and-local Network

We design the GL module (See Figure 2) to model the input feature's global and local information. Given an input tensor $X \in \mathbb{R}^{W \times H \times C}$, local spatial information is encoded by $n \times n$ convolution, and then the feature dimension $X_1 \in \mathbb{R}^{W \times H \times d}$ is extended using point-wise convolution. To model long-range dependence with an effective receptive field of W×H and to let the network learning have a global representation with spatial inductive bias, X_1 is expanded into N flattened patches $X_U \in \mathbb{R}^{P \times N \times d}$,where $P = w \times h$, and $N = W \times H$ is the number of patches, and h and w are the height and width of the patch, respectively. For each $p \in \{1, \ldots, P\}$, the relationship between the patches is encoded by the transformer to obtain $X_G \in \mathbb{R}^{P \times N \times d}$, the GL block captures the local information within each patch and the global information between different patches. Thus, we can fold $X_G \in \mathbb{R}^{P \times N \times d}$ to obtain $X_f \in \mathbb{R}^{W \times H \times d}$. The X_f channel is then projected to a lower C-dimensional space using point-wise convolution and stacked with the input feature X by a cascade operation. Another $n \times n$ convolution layer is then used to fuse the stacked features. Since X_U uses convolution to encode the local information in the $n \times n$ region and X_G encodes the global information of the patch, each pixel X_G can encode the information of all pixels in X with global receptive field. Considering the number of channels to be matched, the network structure should contain two branches, 3×3 convolution and residual mapping, as shown in Figure 2(a). 1×1 convolution can be regarded as a special case of 3×3 convolution (an expanded convolution with edge elements filled with 0). The residual module in the network does not introduce additional computation, but it uses more memory than the single-branch structure. The residual structure makes deeper network models possible, but more branches slow down the model training efficiency. To ensure the stability of the output feature map, a 3×3 convolution with a weight of 1 is initialized. According to the convolutional addition principle, the weights and biases of all branches are further superimposed to obtain the fused 3×3 convolutional layer. We use structural reparameterization to merge the multi-branch structure of the convolutional layer into the single-branch structure in Figure 2(b) to achieve higher inference efficiency[28].

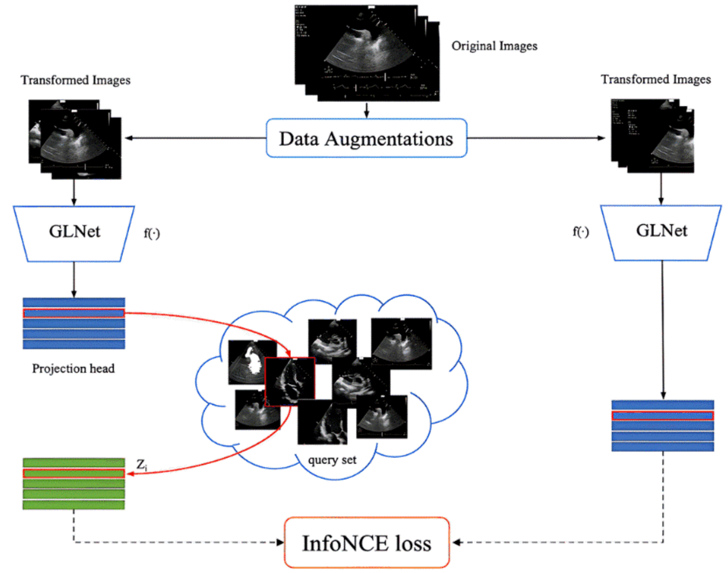


Figure 3. A Query-based Contrastive Learning Representation Framework. QueryCLR uses similar samples of input images as positive pairs in the query set and employs contrastive learning to minimize the distance between positive pairs, aiming to enhance performance.

B. Query Contrastive Learning of Visual Representations

Conventional contrastive learning treats two images obtained from one image by different data augmentations as a positive pair, and $2(N-1)$ (assuming that there are N images in a batch, and after enhancement, they become $2N$ images) images obtained from other $N-1$ images data augmentations negative samples[3]. The representation is learned by maximizing the agreement between 00involves applying various data augmentations to the same image to obtain different perspectives. Data augmentation of the same image cannot provide diverse perspectives of the image, variations of the same object, or other similar entities within the same category.

SimCLR uses two data-augmented embeddings (z_i, z_i^+) (z is the representation obtained after the encoder and the projection head) as positive pairs, with the negative examples coming from the rest of the images in the same batch. Positive pairs from the same image lack variations of the same object[29]. To solve this problem, we find the same image and samples similar to that image as positive pairs in the dataset. Alternatively, a queue Q is used to store more different positive pairs, Q is called a query set. The formula for obtaining similar samples from a query set is as follows:

$$Similar(z, Q) = argmin_{q\in Q}\|z-q\|_2 \qquad (1)$$

The query set Q is updated in a first-in-first-out queue. At the end of each iteration, the embeddings from the training step are placed at the end of the queue, and the first n embeddings are discarded. In addition, it is recommended to ensure that the query set Q is as large as necessary to fit the whole dataset as possible, to approximate the embedding of the whole dataset in vector space. Indeed, Q cannot be infinitely large and we explore the performance of setting differently Q later on. As shown in Figure 3, a picture X is first transformed into X_1 and X_2 by different data augmentation, and X_1 and X_2 are encoded into representations z_1 and z_2 using GLNet and Projection head, noting that the two Encoder and Projection heads share parameters. The Query set maintains a large queue from which the corresponding encodings are taken to provide positive examples for the first set of image augmentation views required by contrastive learning between positive and negative examples. At the same time, the feature representation encoding from the latest batch is put into the queue Q, while the image encoding corresponding to the oldest batch is taken out of the queue so that the content of the encoding in the queue can be continuously updated.

C. Loss Function

In SimCLR, the training direction of the model is guided by drawing close the similarity between z_1 and z_2 and drawing far the similarity between z_1 and $N-1$ negative examples with the following equation:

$$L_z = -\log \frac{\exp(z_1 \cdot z_2/\tau)}{\sum_{k=1}^{N} \exp(z_1 \cdot z_2/\tau)} \qquad (2)$$

Where τ denotes the temperature coefficient. Our positive samples are from query set, so the loss function is optimized as:

$$L_z = -\log \frac{\exp\left(Similar(z_i, Q) \cdot z_i^+/\tau\right)}{\sum_{k=1}^{N} \exp\left(Similar(z_i, Q) \cdot z_i^+/\tau\right)} \qquad (3)$$

Where z_i represents a sample; z_i^+ is a positive sample of the same object as z_i; Q is a query set containing positive samples and multiple negative samples; $Similar()$ calculated using cosine similarity. In this way, the loss function encourages increasing the similarity between the anchor point and the positive sample while decreasing the similarity between the anchor point and the negative sample.

IV. Experiments and Results

A. DataSets

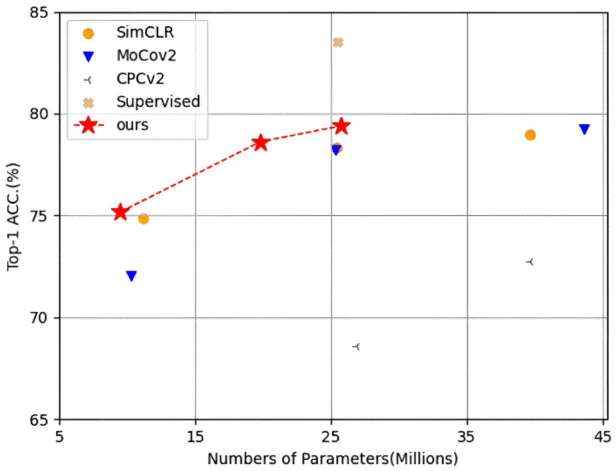


Figure 4. Echocardiogram dataset Top-1 accuracy of linear classifiers trained on representations learned with different self-supervised methods.

We have compiled a dataset of echocardiograms from various populations, consisting of 23,000 images across 12 categories. These categories include the aortic arch, aortic valve, main pulmonary artery, subxiphoid biventricular, subxiphoid four-chambered heart, left ventricular short axis, left ventricular long axis, apical two-chambered heart, apical three-chambered heart, apical four-chambered heart, and apical five-chambered heart. The resolution of the image is 800×600.

B. Training Details

In the process of training our models, we have used a large batch size to encompass a more comprehensive range of samples, thereby enhancing the diversity of the training data. This strategy, however, introduces the challenge of reduced weight adjustments per training iteration. To counteract this, a straightforward approach might be to increase the learning rate. Nonetheless, an increased learning rate at the beginning can cause network divergence because of the random initialization of model weights. To mitigate this issue, we initiate the training regimen with a preliminary 10-epoch run, which allows the model to reach a state of relative stability before applying the predetermined learning rate.

In conjunction with this, we utilize the Layer-wise Adaptive Rate Scaling (LARS) optimizer[30], which tailors the learning rate to each layer, contributing to more stable and effective training dynamics. We set the temperature coefficient at 0.1, begin with an initial learning rate of 0.3, and employ a cosine decay schedule for the learning rate. Furthermore, we incorporate a linear warm-up phase during the initial 10 epochs to expedite model convergence.

For the encoder architecture, we default to using the GLNet-m variant, which produces a 2048-dimensional embedding as the encoder output. The projection head consists of a three-layer Multilayer Perceptron (MLP) with a dimensional sequence of $[2048, 2048, d]$, where d is set to 256. After the pre-training phase on the unlabeled echocardiogram dataset, we removed the projection head and incorporated a fully connected layer to function as a linear classifier. This is followed by fine-tuning the entire network using a limited subset of labels.

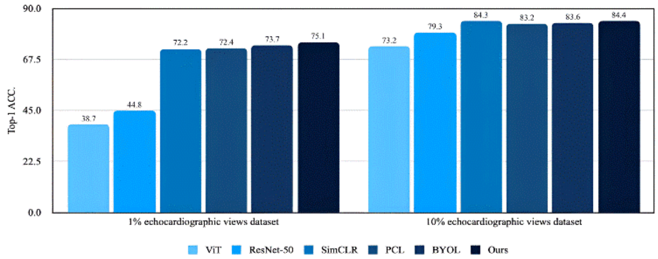


Figure 5. Results of semi-supervised comparison of QueryCLR with other networks. The performance is reported when fine-tuning the pre-trained GLNet with 1% and 10% of the echocardiogram dataset.

The most effective results for QueryCLR were observed when using a queue size of 4096 in combination with a base learning rate of 0.3. All experiments were conducted on a single NVIDIA RTX 3060 GPU to ensure consistent computational resources and reproducible results under standardized hardware conditions

C. Results

As shown in Figure 4, the proposed QueryCLR demonstrated high-quality recognition on echocardiographic datasets compared to supervised ResNet50 and other self-supervised networks.

Table1. presents a direct comparison of various supervised and automatic systems utilized for the echocardiogram dataset, focusing exclusively on detection outcomes. It illustrates the effectiveness of diverse network designs like VGG-16, ResNet-50, InceptionV3, Vision Transformer (ViT), alongside representation learning strategies such as CPCv2, SimCLR, MoCov2, and our proposed GLNet in various sizes (small, medium, large).

Table 1. Detection results of the mainstream supervised and self-supervised on our dataset.

Network	Top-1 ACC. (%)	Top-5 ACC. (%)
Supervised baseline		
VGG	76.2	92.3
ResNet	83.5	98.8
Inception	83.6	99.1
Vision transformer	85.8	99.2
Method using representation learning only		
CPCv2	72.7	78.8
SimCLR	78.3	82.8
SimCLR (w. 10% labels)	84.3	98.2
MoCov2	78.2	82.4
MoCov2 (w. 10% labels)	84.1	89.9
Ours (GLNet-s)	**76.2**	**85.7**
Ours (GLNet-m)	**78.6**	**88.3**
Ours (GLNet-l)	**79.4**	**90.4**
Ours (GLNet-m) (w. 10% labels)	**84.4**	**98.3**

Each method is evaluated based on several parameters: the architecture used, the number of parameters in millions (M), the computational complexity in Giga Floating Point Operations per second (GFLOPs), and their Top-1 and Top-5 accuracy percentages.

Supervised baseline methods such as VGG-16, ResNet-50, InceptionV3, and ViT demonstrate high accuracy, with ViT achieving the highest Top-1 and Top-5 accuracy among them. In contrast, self-supervised methods such as CPCv2, SimCLR, and MoCov2 demonstrate varying degrees of effectiveness, with MoCov2 achieving notable results, especially when trained with only 10% of the labels.

Our proposed GLNet, available in small, medium, and large configurations, demonstrates impressive results. The medium-sized GLNet (GLNet-m) stands out, particularly with a Top-1 accuracy of 78.6% and a Top-5 accuracy of 88.3%, surpassing other self-supervised methods. More importantly, when trained with only 10% of labels, GLNet-m achieves a Top-1 accuracy of 84.4% and a Top-5 accuracy of 98.3%, demonstrating its efficiency and effectiveness in learning from limited labeled data. This highlights the superiority of our model, not only in terms of accuracy but also in its reduced parameter count and computational efficiency, making it a significant contribution to the field of echocardiogram analysis.

We evaluate the validity of our features semi-supervised on 1% and 10% subsets of the echocardiogram dataset following a standard evaluation protocol. We present these results in Figure 5. The first key result of Figure 7 is that our method outperforms all state-of-the-art (SOTA) methods in terms of semi-supervised learning on the 1% subset, demonstrating the strong generalization ability of QueryCLR features. Using a 10% subset, QueryCLR outperforms SimCLR and other methods.

V. Conclusion

This study introduces QueryCLR, a novel framework for contrastive learning of visual representations, designed to enhance the diversity of positive examples by using similar representations from a query set. We developed GLNet, a hybrid CNN-Transformer architecture overcoming CNN's limitations in global representation and Transformers in local information processing, through a diffusion mechanism for transitioning from local to global representations. Additionally, we created a unique echocardiogram dataset for evaluating QueryCLR's effectiveness and advancing academic research. Our findings show QueryCLR surpasses other self-supervised methods, achieving a top-1 accuracy of 79.4%, and outperforms current state-of-the-art methods in semi-supervised training scenarios. Future work will extend its application to various downstream tasks, including segmentation and object detection, to further explore and utilize its capabilities.

Acknowledgment

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] Andre *et al.*, "A guide to deep learning in healthcare," 2019.

[2] L. Dai *et al.*, "A deep learning system for detecting diabetic retinopathy across the disease spectrum," vol. 12, no. 1, p. 3242, 2021.

[3] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A Simple Framework for Contrastive Learning of Visual Representations," 2020.

[4] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729-9738.

[5] S. Bae, S. Kim, J. Ko, G. Lee, S. Noh, and S.-Y. Yun, "Self-Contrastive Learning: An Efficient Supervised Contrastive Framework with Single-view and Sub-network," 2022.

[6] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, "A Survey on Contrastive Self-Supervised Learning," *Technologies*, vol. 9, no. 1, p. 2, 2021.

[7] K. O'Shea and R. Nash, "An Introduction to Convolutional Neural Networks," *ArXiv e-prints*, 11/01 2015.

[8] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *ArXiv*, vol. abs/2010.11929, 2020.

[9] A. Madani, J. R. Ong, A. Tibrewal, and M. A.-O. Mofrad, "Deep echocardiography: data-efficient supervised and semi-supervised deep learning towards the automated diagnosis of cardiac disease," (in eng), no. 2398-6352 (Electronic).

[10] J. Zhang *et al.*, "Fully Automated Echocardiogram Interpretation in Clinical Practice," (in eng), no. 1524-4539 (Electronic).

[11] X. Gao, W. Li, M. Loomes, and L. Wang, "A fused deep learning architecture for viewpoint classification of echocardiography," *Information Fusion*, vol. 36, pp. 103-113, 2017/07/01/ 2017.

[12] A. Ghorbani *et al.*, "Deep learning interpretation of echocardiograms," *npj Digital Medicine*, vol. 3, no. 1, p. 10, 2020/01/24 2020.

[13] Y. Deng *et al.*, "Myocardial strain analysis of echocardiography based on deep learning," (in eng), no. 2297-055X (Print).

[14] M. Wilson Pfr Fau - Gilany *et al.*, "Self-Supervised Learning With Limited Labeled Data for Prostate Cancer Detection in High-Frequency Ultrasound," (in eng), no. 1525-8955 (Electronic).

[15] B. VanBerlo, B. Li, A. Wong, J. Hoey, and R. Arntfield, "Exploring the Utility of Self-Supervised Pretraining Strategies for the Detection of Absent Lung Sliding in M-Mode Lung Ultrasound," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023, pp. 3077-3086.

[16] K. He, X. Zhang, S. Ren, and J. J. I. Sun, "Deep Residual Learning for Image Recognition," 2016.

[17] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," 2015.

[18] D. Dwibedi, Y. Aytar, J. Tompson, P. Sermanet, and A. Zisserman, "With a Little Help from My Friends: Nearest-Neighbor Contrastive Learning of Visual Representations," 2021.

[19] Z. Xie *et al.*, "Simmim: A simple framework for masked image modeling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9653-9663.

[20] Z. Qin, P. Zhang, F. Wu, and X. Li, "Fcanet: Frequency channel attention networks," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 783-792.

[21] G. Tang, M. Müller, A. Rios, and R. Sennrich, *Why Self-Attention? A Targeted Evaluation of Neural Machine Translation Architectures*. 2018.

[22] J. Sinha A Fau - Dolz and J. Dolz, "Multi-Scale Self-Guided Attention for Medical Image Segmentation," (in eng), no. 2168-2208 (Electronic).

[23] J. Hu, L. Shen, S. Albanie, G. Sun, E. J. I. t. o. p. a. Wu, and m. intelligence, "Squeeze-and-Excitation Networks," vol. 42, no. 8, pp. 2011-2023, 2020.

[24] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3-19.

[25] Y. Chen, X. Dai, D. Chen, M. Liu, and Z. Liu, "Mobile-Former: Bridging MobileNet and Transformer," 2021.

[26] A. G. Howard *et al.*, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," 2017.

[27] A. Srinivas, T. Y. Lin, N. Parmar, J. Shlens, and A. Vaswani, "Bottleneck Transformers for Visual Recognition," 2021.

[28] X. Ding, X. Zhang, N. Ma, J. Han, G. Ding, and J. Sun, "Repvgg: Making vgg-style convnets great again," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 13733-13742.

[29] R. Zhu, B. Zhao, J. Liu, Z. Sun, and C. Chen, *Improving Contrastive Learning by Visualizing Feature Transformation*. 2021.

[30] Y. You, I. Gitman, and B. Ginsburg, "Large Batch Training of Convolutional Networks," ed, 2017.