









RESEARCH ARTICLE

WILEY

GLRP: Global and local contrastive learning based on relative position for medical image segmentation on cardiac MRI

Xin Zhao¹  | Tongming Wang¹  | Jingsong Chen¹  | Bingrun Jiang¹  |
Haotian Li¹  | Nan Zhang²  | Guang Yang^{3,4,5,6}  | Senchun Chai¹ 

¹School of Automation, Beijing Institute of Technology, Beijing, China

²Department of Radiology, Beijing Anzhen Hospital, Capital Medical University, Beijing, China

³National Heart and Lung Institute, Imperial College London, London, UK

⁴Bioengineering Department and Imperial-X, Imperial College London, London, UK

⁵Cardiovascular Research Centre, Royal Brompton Hospital, London, UK

⁶School of Biomedical Engineering & Imaging Sciences, King's College London, London, UK

Correspondence

Guang Yang, Bioengineering Department and Imperial-X, Imperial College London, London W12 7SL, UK.

Email: g.yang@imperial.ac.uk

Senchun Chai, School of Automation, Beijing Institute of Technology, No. 5 Zhongguancun South Street, Haidian District, Beijing 100081, China.
Email: chaisc97@163.com

Funding information

ERC IMI, Grant/Award Number: 101005122; H2020, Grant/Award Number: 952172; Medical Research Council (MRC), Grant/Award Number: MC/PC/21013; Royal Society, Grant/Award Number: IEC \NSFC\211235; NVIDIA Academic Hardware Grant Program; Boehringer Ingelheim Ltd; UKRI Future Leaders Fellowship, Grant/Award Number: MR/V023799/1

Abstract

Contrastive learning, as an unsupervised technique, is widely employed in image segmentation to enhance segmentation performance even when working with small labeled datasets. However, generating positive and negative data pairs for medical image segmentation poses a challenge due to the presence of similar tissues and organs across different slices in datasets. To tackle this issue, we propose a novel contrastive learning strategy that leverages the relative position differences between image slices. Additionally, we combine global and local features to address this problem effectively. In order to enhance segmentation accuracy and reduce isolated mis-segmented regions, we employ a two-dimensional fully connected conditional random field for iterative optimization of the segmentation results. With only 10 labeled samples, our proposed method is able to achieve average dice scores of 0.876 and 0.899 on the public and private dataset heart segmentation tasks, surpassing the PCL method's 0.801 and 0.852. Experimental results on both public and private MRI datasets demonstrate that our proposed method yields significant improvements in medical segmentation tasks with limited annotated samples, outperforming existing semi-supervised and self-supervised techniques.

KEYWORDS

contrastive learning, medical image segmentation, relative position

1 | INTRODUCTION

In the diagnosis of myocardial disease, magnetic resonance imaging (MRI) is widely used for non-invasive assessment of cardiac structure and function due to its superior image resolution and soft tissue contrast.^{1,2} Segmentation is the first step in diagnosis. In recent years, fully convolutional neural network (FCN),³ U-net network,⁴ and nn U-net⁵ have achieved great success in cardiac MRI segmentation due to their multi-scale feature extraction and fusion capabilities.⁶ After U-net was proposed, some researchers have proposed modules such as residual connectivity,⁷ dense connectivity,⁸ and attention mechanism^{9,10} to further improve the original model. In recent years, with the hot research of transformer in the image field,¹¹ the transformer network-based medical image segmentation model represented by Trans-Unet¹² and Swin-Unet¹³ has become the most advanced and cutting-edge method. However, most of the existing supervised learning methods cannot achieve satisfactory results in the absence of rich labeled data. They are effective only after the completion of extensive annotation efforts. In addition, there are still challenges such as generalization errors, false correlations, and adversarial attacks¹⁴ when applying supervised learning methods. Therefore, many researchers have proposed alternative methods when annotation samples are scarce or the original image quality is poor. Sun et al. proposed Meta self-Attention Prototype Incrementer (MAPIC) to achieve Few-shot class-incremental classification in medical time series.¹⁵ Wang et al. proposed a novel image enhancement scheme to make low-brightness images clear and natural.¹⁶ Some researchers focus on biological tissue modeling,¹⁷ feature tracking,¹⁸ and image matching,¹⁹ and their proposed methods can combine medical prior knowledge to support typical problems such as segmentation in the field of medical image analysis. Additionally, many researchers also use unsupervised or self-supervised learning methods as a better alternative to supervised learning. As an unsupervised learning method, the Stacked Sparse Autoencoder (SSAE) proposed by Qadri et al. can extract discriminative features from unlabeled patches for CT vertebral segmentation.²⁰ The Pa-DBN-BC model proposed by Hirra et al. automatically extracts features from patches and then uses logistic regression to classify pathological images.²¹ Their study was able to make better use of the information from the images themselves, and also took into account the effects of overlapping structures, ambiguous object boundaries, and image variations. In the SSL setting, proxy supervision information derived from the image itself is used to construct a pretext task to train the model. Then, the model learns the features of the images

for downstream segmentation tasks to achieve segmentation accuracy comparable to that of conventional supervised learning methods but using fewer labeled datasets.

Contrastive learning is a popular and successful variant of SSL. A basic implementation of contrastive learning is to construct pairs of positive and negative samples by transforming them into different representations. A neural network is then trained to cluster the representations of similar pairs together and separate the representations of dissimilar pairs by minimizing a contrastive loss defined by the researchers. During this training process, the model will ignore the nuance of the sample images and learn the quality representations of the sample images. In practice, the model is used as a good initialization for downstream supervised segmentation tasks.²²

Although contrastive learning has achieved great success in the field of medical image segmentation, we believe that a limitation of previous work in the construction of sample pairs hinders better segmentation performance. In particular, the various contrastive learning strategies proposed so far have been too inflexible in defining positive and negative sample pairs. Zeng et al.²³ proposed to generate contrastive data pairs by comparing the difference in slice position with a given threshold. Chaitanya et al.²⁴ divided the three-dimensional medical image into several regions, and slices in different regions were mutually negative sample pairs, while slices in the same region were positive sample pairs. These construction methods for generating comparison samples have achieved success on many public datasets.

However, these proposed methods make any two samples necessarily positive or negative sample pairs. In other words, they cannot measure the similarity between these data pairs, which lacks rationality. In contrast, this study discards the concept of positive and negative samples and proposes a strategy based on relative position to define the similarity of each sample pair. We perform experiments on private dataset and public Automated Cardiac Diagnosis Challenge (ACDC) dataset and compare other existing contrastive learning methods with the proposed method. The results show that the method in this paper achieves superior segmentation performance.

Our main contributions are as follows:

1. We proposed a global contrastive learning strategy based on relative position. This approach can use the relationship between slice similarity and relative position in 3D medical images to guide the network to learn high-quality feature representations of the images, improve pre-training performance, and avoid potential misclassification of positive and negative

samples by previous positional contrast learning methods

2. We proposed a local contrastive learning strategy based on relative position. This method can complete the pre-training by exploiting the similarity between patches within a slice, thus effectively exploiting the relative position relationship between patches, which can further improve the segmentation performance when combined with the global contrastive learning strategy
3. We used a fully connected CRF to iteratively optimize the mislabeled segmentation results. It can process the segmentation results by considering the relationship between all pixels in the original image, thus correcting isolated mis-segmented regions while obtaining finer segmentation boundaries

2 | RELATED WORK

In recent years, there have been many applications of contrastive learning in the field of image segmentation. Wang et al.²⁵ proposed a supervised semantic segmentation approach based on pixel-level contrastive learning, which focuses on the dependency relationship between pixels. This method considers the global semantic similarity of all pixels in the whole training set and uses more diverse and large samples to obtain a better semantic feature space.

In the field of medical image segmentation, researchers mainly focus on the construction of contrastive sample pairs. You et al.²⁶ proposed to construct the contrastive loss based on the signed distance map (SDM), the output of the teacher network model, to ensure that the model can extract boundary knowledge. Liu et al.²⁷ associated local patches from the input space with hidden vectors in the output feature map of the encoder network to construct patch-level positive and negative sample pairs. Chaitanya et al.²⁴ proposed contrastive learning based on global and local features, which spatially divides the cardiac short-axis MRI images into four regions, and the slices within the same region are considered as positive samples, and the slices on different regions are considered as negative samples. In addition, the authors also propose a contrastive learning strategy for local features with the goal of consistent local representations of images of the same type and dividing the feature map of the layer L decoder into N image patches, and the images defined as positive samples in terms of global features, whose image patches at the same position are only considered as positive samples from each other, otherwise they are considered as negative samples from each other.

Chaitanya et al.²⁸ proposed a local contrast loss combining unlabeled images and limited annotated images and performed pixel-level contrast learning training based on pseudo-labeled pixel categories to learn good pixel-level features useful for segmentation. In References 23,28,29, researchers designed sampling strategies based on the structural similarity images to complete the definition of positive and negative sample pairs, which achieved satisfactory results.

3 | METHOD

3.1 | GLRP segmentation framework

We used the idea of self-supervised learning to pre-train the segmentation model before fine-tuning it. In the pre-training stage, a pretext task is defined to pre-train the network model on many unlabeled samples to learn the representation of the downstream tasks. In order to compute subsequent contrastive losses, the output of the network coder is augmented with a projection head that maps high-dimensional feature representations to low-dimensional feature vectors. The projection head is used only in the pre-training phase and removed in the fine-tuning phase, thus avoiding additional computational costs in the segmentation task.²⁵ The network model is then fine-tuned for specific downstream tasks using labeled samples. Finally, the prediction results are processed by the 2D fully connected CRF model³⁰ to obtain the final prediction result.

In this study, we choose 2D U-net as the backbone model for image segmentation. This network is a commonly used structure in the field of medical image segmentation, which can achieve satisfactory results in most segmentation tasks. During the self-supervised pre-training stage, the network takes an unlabeled image slice of size 288×288 as input, and it uses both the position information between slices and the position information within a slice as labels. As shown in Figure 1, after five encoder blocks with convolution and max pooling layers and a projection head with an average pooling layer and two fully connected layers, the image is transformed into a 1×512 vector as output.

For the supervised fine-tuning stage, the input is a labeled image slice of size 288×288 . The label is the ground truth of the slice, and the output is the predicted segmentation result of the network. In the multi-model fusion stage, the input is the prediction results under three different contrastive learning strategies, and the output is the final segmentation result. The entire self-supervised segmentation framework of this study is shown in Figure 2.

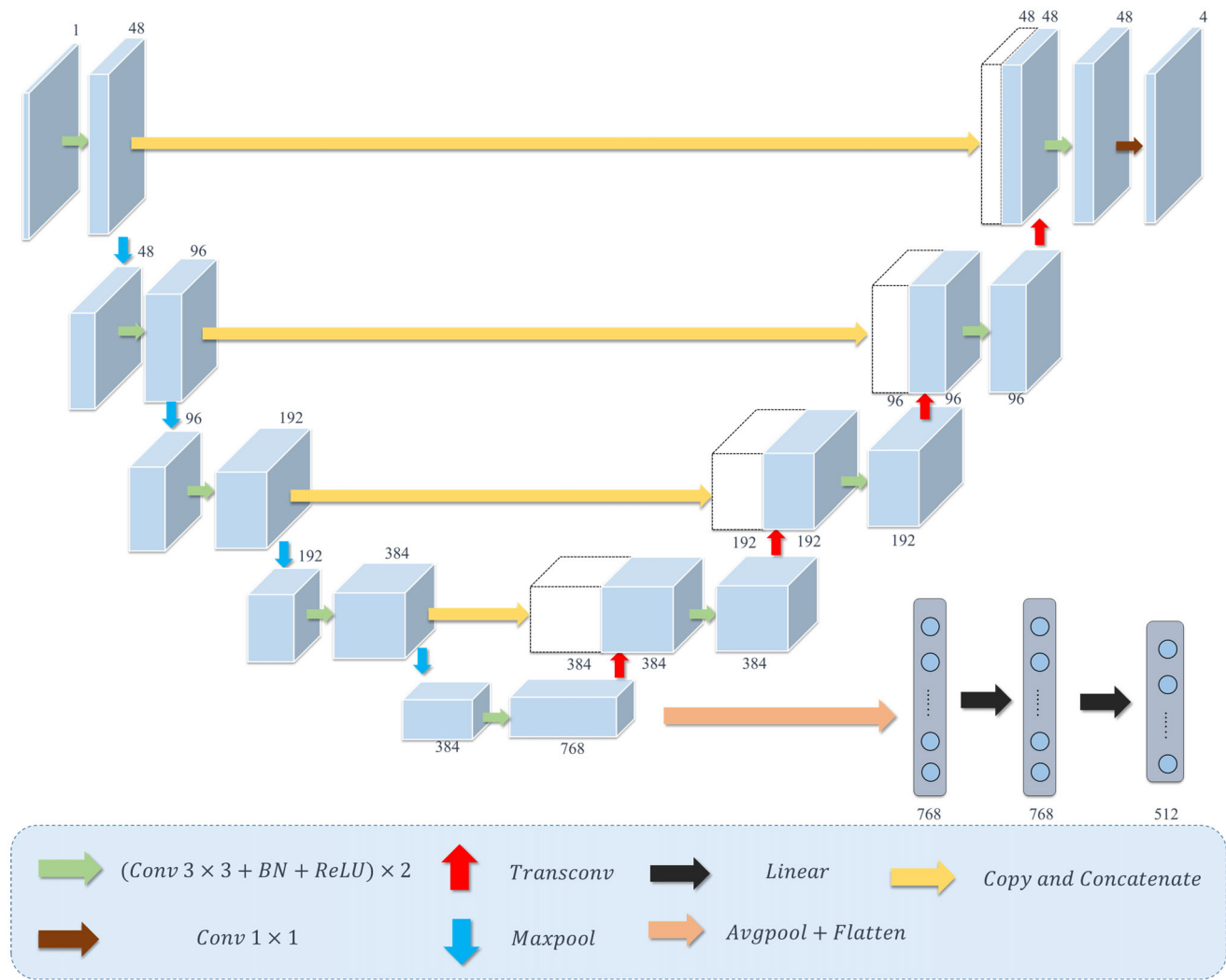


FIGURE 1 The overall structure of the network. Where $(\text{Conv } 3 \times 3 + \text{BN} + \text{ReLU}) \times 2$ represents two successive convolution modules, each containing a convolution layer, a batch normalization (BN), and a ReLU activation function, the convolution kernel of the convolution layer has a size of 3×3 , the number of filters is 48, and the step size is 1. Maxpool is the maximum pooling layer with a pooling window of 2. Transconv is a transposed convolution layer. Avgpool is an average pooling layer, and Linear is a fully connected layer.

3.2 | Data pre-processing

Due to differences in image size, spatial resolution, and so forth between different patients, it is necessary to pre-process the data and eliminate the differences between different images to make the processed data more suitable for network training. Preprocessing mainly includes N4 bias field correction, image resampling and scaling, data standardization, and data augmentation.

3.2.1 | Data augmentation

We used data augmentation to obtain $2N$ data samples with N unlabeled data samples in the training set. The specific data augmentation operations used include

random rotation, random flipping, random elastic deformation, and random scaling.

3.3 | Self-supervised pre-training stage

3.3.1 | Self-supervised label definition

After data preprocessing, a position label was generated for each slice as a self-supervised signal.²³ The position label was defined as the relative position of the slice along the cardiac short axis in the heart. Suppose the number of the cardiac short-axis MRI slices is m . Then we define the position label of the n th slice along the direction from the base of the heart to the apex as n/m . The schematic diagram of the slice position label definition is shown in Figure 3.

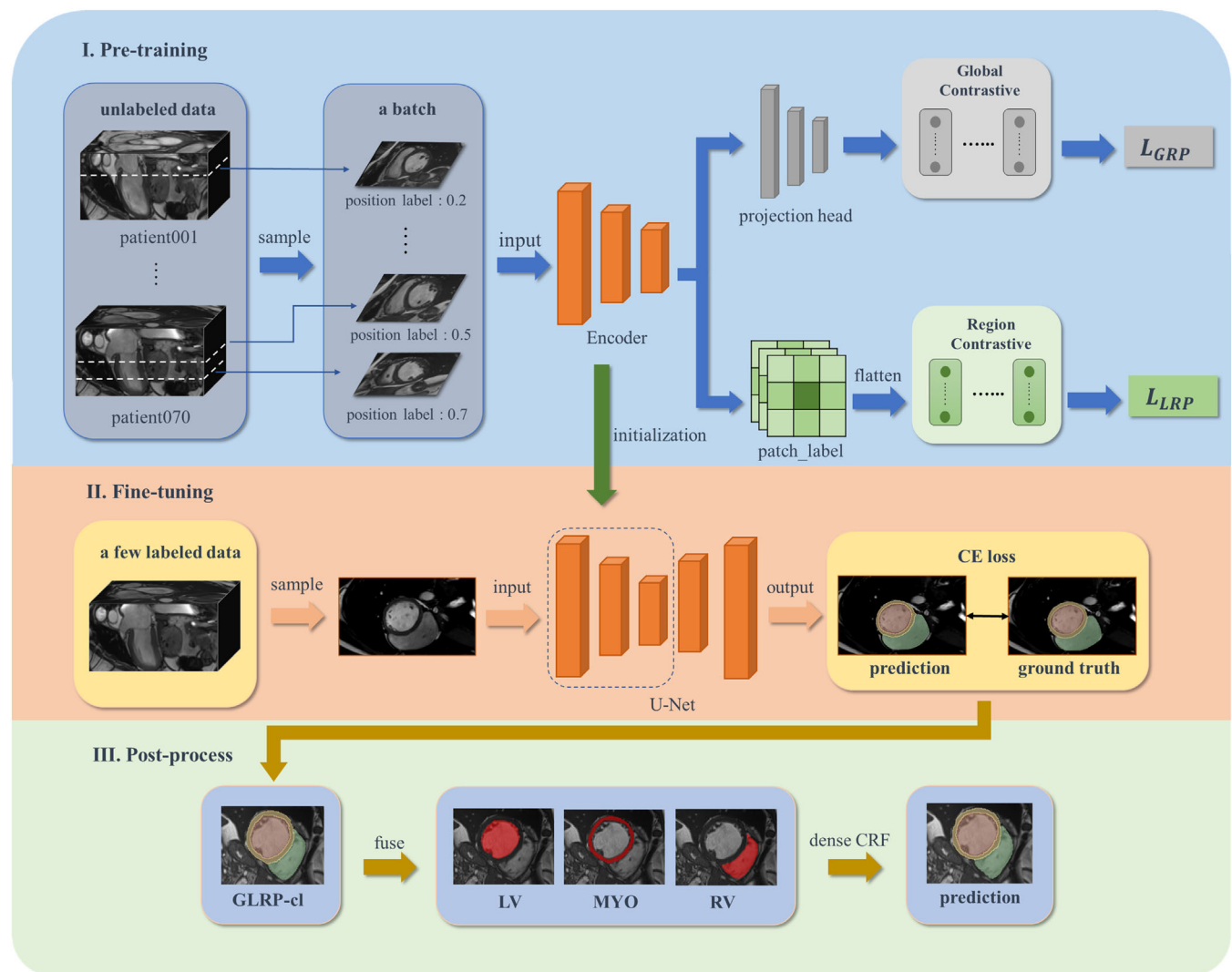
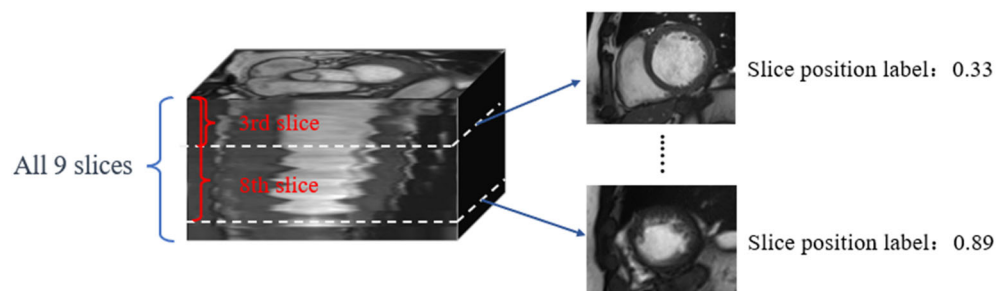


FIGURE 2 The proposed GLRP segmentation framework consists of three components: (A) a pre-training part based on a global and local contrastive learning strategy, (B) a supervised fine-tuning part, and (C) a post-process part based on a 2D fully connected CRF model.

FIGURE 3 The definition of the slice position label.



3.3.2 | Pretext task definition

In this study, we used relative position-based global contrast learning and local contrastive learning as pretext tasks to pre-train the network. Therefore, the encoder can learn the similarities and differences of the sample

pairs, and then obtain the high-quality encoding and feature representation of the images.

Global contrastive learning strategy

In a medical image, the slice-level image information is often similar at the same or adjacent slice levels of images

from the same patient or different patients, and this similarity decreases sharply as the slice distance increases.

Inspired by this,²³ proposed position contrastive loss (PCL). This strategy defines positive and negative sample pairs according to whether the distance difference between the image slices is within the threshold. However, the adjustment of the threshold parameters is labor intensive. Furthermore, it is not reasonable to consider slices above a threshold as negative sample pairs, because medical images are collections of different dimensions of a single organ, and therefore there is always some similarity between slices. Therefore, we use the relative position information instead of the threshold as the self-monitoring signal, and we do not construct positive and negative sample pairs. The global similarity between two slices based on the relative position information can be defined as follows:

$$d_{ij} = \frac{1}{1 + a(x_i - x_j)^2} \quad (1)$$

where d_{ij} is the global similarity of slice i and j . x_i and x_j are the position label of slices i and j . a is an adjustable hyperparameter called similarity coefficient which reflects the similarity between sample pairs, and we find that setting it to 15 works best after several trials. In cardiac short-axis MRI images, the image similarity decreases sharply as the distance difference increases. Therefore, the position difference between the slices is in the denominator. If x_i and x_j are the same slice, the global similarity is 1.

After we obtained the samples from the pre-training stage and their corresponding labels. We began to train the encoder and projection head of the network. The contrastive loss function³¹ L_{GRP} is:

$$L_{GRP} = \sum_{i=1}^{2N} -\frac{1}{2N-1} \sum_{j=1, j \neq i}^{2N-1} \log \frac{d_{ij} e^{\frac{\text{sim}(z_i, z_j)}{\tau}}}{\sum_{k=1, k \neq i}^{2N-1} d_{ik} e^{\frac{\text{sim}(z_i, z_k)}{\tau}}} \quad (2)$$

where z_i is a learned representation vector of slice i . τ is a temperature parameter to adjust the loss value. $\text{sim}(z_i, z_j)$ is the cosine similarity whose calculation result is used to measure the similarity of two representation vectors in space. The cosine similarity can be calculated by:

$$\text{sim}(z_i, z_j) = \frac{z_i \cdot z_j}{\|z_i\| \|z_j\|} \quad (3)$$

Local contrastive learning strategy

The global contrastive learning strategy only considers the positional information at the slice level, but there are

also relative positional relationships within the slice. To explore the internal correlation of the encoder output feature map and to better guide the encoder to learn high-order features of images, this section associates the feature map with the original image based on the concept of the receptive field (RF).

Specifically, in this section, the relative position definition method divides the feature map obtained by the encoder for each slice into nine image patches. Each image patch is obtained by layer-by-layer convolution and down-sampling of a certain area of the original image, so the receptive field means that all pixels of the feature map patch correspond to the size of the original image area. By finding the receptive field of each image patch on the original image and calculating the SSIM³² of the receptive field regions of the two image patches, the similarity matrix of the image patches can be obtained to measure the relative positive and negative sample relationships between the image patches.

Considering the inconsistent orientation of cardiac short-axis MRI images due to scan angle differences and missing affine matrix, as well as operations such as random rotation and flipping performed on the dataset during data augmentation, the local disparity of the images is mainly reflected in the distance of each patch from the image center. Therefore, when computing the similarity matrix of the image patches, we assume that the image patches in the four corner positions of the nine image patches are equivalent, and the four image patches adjacent to the central image patch are equivalent.

Based on the above ideas, this section obtains the similarity matrix of feature map patches based on the similarity of receptive fields and constructs corresponding position labels based on the relative position of feature map patches. In summary, the framework of the local contrastive learning strategy is shown in Figure 4.

Combining the position labels of the image patches with the slice position labels where the image patches are located, based on the InfoNCE contrastive loss function, we propose to compute the local contrastive loss based on the position information of the image patches:

$$L_{LRP} = \sum_{l=1}^{18N} -\frac{1}{18N-1} \sum_{m=1, m \neq l}^{18N-1} \times \log \frac{\text{PSM}(l, m) \cdot d_{ij} \cdot e^{\frac{\text{sim}(z_l, z_m)}{\tau}}}{\sum_{n=1, n \neq l}^{18N-1} \text{PSM}(l, n) \cdot d_{ik} \cdot e^{\frac{\text{sim}(z_l, z_n)}{\tau}}} \quad (4)$$

where i, j, k represent the slices in which the image patches l, m, n are located, and PSM is the similarity matrix of the image patches.

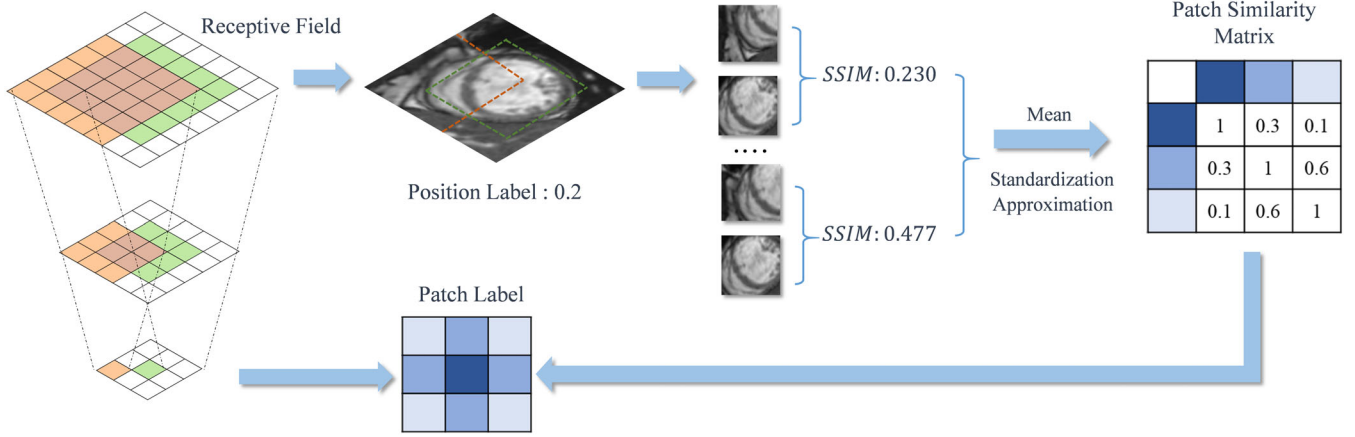


FIGURE 4 Framework of local contrastive learning strategy based on region perception.

In the model pre-training stage, we combine the loss function of the global contrastive learning strategy and local contrastive learning strategy. Then we take their linear combination as the final total loss function:

In the pre-training phase of the model, we combine the loss functions of the global contrast learning strategy and the local contrast learning strategy. We then use a linear combination of them as the overall loss function:

$$L = \lambda_1 L_{GRP} + \lambda_2 L_{LRP} \quad (5)$$

where λ_1 and λ_2 are the weight coefficients of the global and local strategies, respectively. When we use only the global contrastive learning strategy, λ_1 is set to 1 and λ_2 is set to 0. When we use only the local contrastive learning strategy, λ_1 is set to 0 and λ_2 is set to 1. When both strategies are used, we find that the best results are obtained when both λ_1 and λ_2 are set to 0.5 after the experiments.

3.4 | The fine-tuning stage for supervised segmentation tasks

In the pre-training stage of self-supervised contrastive learning, we have completed the pre-training of the U-Net network encoder. Then, we use the encoder parameters as the initialization parameters of the current model, discard the projection head from the previous stage, and randomly initialize the decoder part of U-Net. We use a small number of labeled datasets in the training set to fine-tune the U-Net segmentation model and finally enable our segmentation network to achieve high segmentation accuracy by learning only a few labeled samples.

At this stage, the loss function is the cross-entropy loss function, which is commonly used in the field of medical image segmentation, which can evaluate the

cross-entropy between the predicted pixel value and the labeled pixel value and make the predicted value close to the label value at the pixel level by minimizing the entropy. The cross-entropy loss function is shown below:

$$CE_{loss}(y_p, y_l) = -\frac{1}{N} \sum_N y_p \log(y_l) + (1 - y_l) \log(1 - y_p) \quad (6)$$

where N is the number of pixel categories, y_p is the predicted segmentation result of the model and y_l is the ground truth.

3.5 | The post-process stage for segmentation tasks

In the pre-training stage, we can obtain three pre-training networks by changing the weights of global and local contrast losses. In the fine-tuning stage, we fine-tune the encoders of the three networks. In our experiments, we found that for irregularly shaped parts, the boundaries of different classes of segmentation results are prone to fragment mis-segmentation regions. The pixels in these mis-segmented regions are unstable, and there are fewer pixels with similar gray values around them. The fully connected CRF optimizes the predicted values of the unstable pixels in the boundary regions by analyzing the similarity between the pixels and their surrounding pixels to achieve the effect of correcting the mis-segmented regions. First, the initial Probability matrix is shown as follows:

$$Q = [Q_0 \quad Q_1 \quad Q_2 \quad Q_3] \quad (7)$$

$$Q_L = q_0 + q_L \times M_L \quad (8)$$

Q_L is a submatrix of Q , and L represents the three split foreground and background categories. Where q_0 is

ALGORITHM 1 Fully connected CRF

Require: Point Set V , Initial Probability Matrix Q

Ensure: X_{new} { Results of iterative optimization }

```

1. for  $k$  times do
2.   for pixel  $x_i$  in  $V$  do
3.     for matrix  $Q_L$  in  $Q$  do
4.       for segmentation class  $l$  do
5.         for smooting kerbel  $m$  do
6.            $Q_{x_i}^m(l) = \sum_{x_j \neq x_i} [Q_{x_j}(l) \cdot k^m(f_{x_i}, f_{x_j})]$ 
7.         end for
8.          $\tilde{Q}_{x_i}(l) = \sum_m \omega^{(m)} Q_{x_i}^m(l)$ 
9.       end for
10.       $\hat{Q}_{x_i}(L) = \sum_{l \in Labels} \mu(L, l) \tilde{Q}_{x_i}(l)$ 
11.       $\hat{Q}_{x_i}(L) = \exp\{-\varphi_u(L) - \hat{Q}_{x_i}(L)\}$ 
12.    end for
13.    for matrix  $Q_L$  IN  $Q$  do
14.       $Q_{x_i}(L) = \frac{\hat{Q}_{x_i}(L)}{\prod_{L \in Labels} Q_{x_i}(L)}$ 
15.    end for
16.     $X_{x_{inew}} = \arg \max_{k \in Labels} Q(k)$ 
17.  end for
18.   $X_{new} = \{L_{x_{inew}}\}$ 
19. end for

```

the weight coefficient and M_L is the model prediction matrix corresponding to label L . The subsequent optimization operations are shown in the CRF algorithm in Algorithm 1.

4 | EXPERIMENTS AND RESULTS

4.1 | Datasets

This study uses a public dataset and a private dataset to evaluate our proposed method. The public dataset is the ACDC (The Automated Cardiac Diagnosis Challenge) dataset.³³ In these images which cover the entire cardiac cycle, only the left ventricle (LV), myocardium (MYO), and right ventricle (RV) of the ES and ED frames are labeled by experienced clinical medical experts.

The private dataset for this study was obtained from Beijing Anzhen Hospital. The dataset contains cardiac MRI short-axis images of 214 patients, and informed consent was waived because of a retrospective study. The cardiac MRI images cover the entire cardiac cycle, ranging from 25 frames to 40 frames for different patients. In addition, the slice image resolutions are 0.87×0.87 –

$2.08 \times 2.08 \text{ mm}^2$ and the image slice thickness is 7–10 mm with no gap between the slices. The LV, MYO, and RV of the ES and ED frames are manually annotated by experienced radiologists using CVI42 software (v5.14.2, Canada).

Both datasets were divided into training, validation, and testing sets in a ratio of 7:1:2 based on the number of patients. It should be noted that in the pre-training stage, all unlabeled data in the training set are used for model training, while in the fine-tuning stage, only the labeled ES and ED frame images in the training set are selected for model fine-tuning, and model testing and validation are performed only in the fine-tuning stage based on segmentation tasks.

4.2 | Pre-processing

For the two datasets shown above, we perform the processing mentioned in 3.2. First, we resample each original image to $1.25 \times 1.25 \text{ mm}^2$ and crop it uniformly to a size of 288×288 . Then, we standardize the image intensity according to the following formula:

$$P_i' = (P_i - \mu) / \sigma \quad (9)$$

where P_i is the pixel value of pixel point i , P_i' is the standardized pixel value, and μ and σ are the mean and standard deviation of the image pixels. Finally, we obtained the dataset that can be used for training through data augmentation.

4.3 | Evaluation

We evaluate the accuracy of the model segmentation results from two aspects: segmentation area and segmentation contour. Specifically, two segmentation indicators, the Dice coefficient and Hausdorff distance (HD), were used.

The closer the Dice coefficient is to 1, the higher the degree of agreement between the predicted label of the model and the true label, and the better the segmentation result. The HD can be expressed as the maximum value of the minimum distance from a point in the region to the contour of another region. Therefore, the smaller the HD, the better the segmentation result.

4.4 | Experiment details

We train and evaluate our method on public and private datasets. All experiments were run on a Tesla V100S PCIe GPU with 32 GB of memory. Initially, a consistent batch

size of 12 was maintained for 200 epochs during both the pre-training and fine-tuning stages. During pre-training, we use SGD as an optimizer with a learning rate of 0.1 and implement a cosine learning rate scheduler. The temperature coefficient (τ) is set to 0.1. Moving to the fine-tuning stage, we used the U-Net architecture and optimized it using cross-entropy loss with a learning rate of 0.0005. Adam served as the optimizer, with the cosine scheduler remaining in use, all based on previous work.²³ However, when applied to our private dataset, the results indicated that the model did not fully converge during the fine-tuning stage. In refining the approach, we increased the batch size, setting it to 30, and adjusted the fine-tuning learning rate from 0.0005 to 0.0001. In addition, both the pre-training and fine-tuning stages were extended to 300 epochs, which contributed to more efficient training and improved model convergence.

In our study, we conducted the following experiments to verify the effectiveness of the proposed strategy. All the above experiments used the Dice coefficient and HD as evaluation indicators and were conducted on the ACDC dataset and the private dataset, where the ROI region of the ACDC dataset is LV, MYO, and RV, and the ROI region of the private dataset is LV and MYO.

4.4.1 | Comparative experiments with other methods

To verify the superiority of Global and local contrastive learning based on relative position (GLRP)-cl, this part

of the experiment compares the results of the randomly initialized, Simclr,³⁴ GCL,²⁴ and PCL²³ methods with the proposed GLRP-cl. The comparison is made from three aspects: (1) no pre-training (Random); (2) using contrastive learning pre-training strategies in the natural image domain (Simclr); and (3) using contrastive learning pre-training strategies in the medical image domain (GCL, PCL) for a comprehensive analysis and comparison. The experiment is based on the pre-training model obtained by GLRP-cl, and then k ($k = 2 \setminus 6 \setminus 10 \setminus 15 \setminus 20 \setminus 30 \setminus 70 \setminus 105$) labeled samples are used in the fine-tuning stage to complete the training of the segmentation model. The experimental results in the form of mean (standard deviation) on the ACDC dataset and the private dataset are shown in Tables 1 and 2, respectively.

The comparison of segmentation effects is shown in Figures 5 and 6, where green annotations represent the RV, yellow annotations represent the MYO, red annotations represent the LV, and the red box indicates the apparent segmentation abnormal area.

The results show that using self-supervised contrastive learning strategies in the natural image domain increases false negative samples and may harm pre-trained models by not providing any gain to the segmentation task; in contrast, the GLRP-cl contrastive learning strategy measures similarity based on the relative position and achieved the highest segmentation accuracy in the comparison experiment, confirming the advanced nature of the proposed GLRP-cl strategy.

TABLE 1 Dice score(mean) of the segmentation results for the ACDC dataset.

Method	$k = 2$	$k = 6$	$k = 10$	$k = 15$	$k = 20$	$k = 30$	$k = 105$
Random	0.421 (0.09)	0.695 (0.09)	0.758 (0.08)	0.834 (0.05)	0.844 (0.06)	0.873 (0.04)	0.906 (0.03)
SimCLR	0.260 (0.14)	0.695 (0.10)	0.771 (0.07)	0.829 (0.05)	0.839 (0.06)	0.877 (0.05)	0.910 (0.02)
GCL	0.310 (0.06)	0.616 (0.15)	0.647 (0.09)	0.772 (0.10)	0.812 (0.06)	0.854 (0.05)	0.911 (0.03)
PCL	0.568 (0.17)	0.746 (0.10)	0.801 (0.07)	0.844 (0.05)	0.847 (0.07)	0.879 (0.04)	0.912 (0.03)
GLRP-cl	0.684 (0.15)	0.816 (0.09)	0.851 (0.06)	0.864 (0.04)	0.876 (0.05)	0.888 (0.03)	0.912 (0.03)

Note: Bold value represents the best dice scores among different methods.

TABLE 2 Dice score of the segmentation results for the private dataset.

Method	$k = 2$	$k = 6$	$k = 10$	$k = 15$	$k = 20$	$k = 30$	$k = 70$
Random	0.065 (0.10)	0.651 (0.16)	0.791 (0.11)	0.807 (0.13)	0.877 (0.06)	0.896 (0.05)	0.922 (0.03)
SimCLR	0.028 (0.04)	0.486 (0.26)	0.743 (0.12)	0.834 (0.07)	0.857 (0.06)	0.885 (0.04)	0.913 (0.04)
GCL	0.365 (0.23)	0.798 (0.07)	0.856 (0.06)	0.887 (0.05)	0.897 (0.04)	0.909 (0.04)	0.921 (0.04)
PCL	0.402 (0.18)	0.789 (0.08)	0.852 (0.07)	0.899 (0.04)	0.903 (0.04)	0.907 (0.04)	0.919 (0.03)
GLRP-cl	0.421 (0.33)	0.804 (0.09)	0.870 (0.06)	0.904 (0.04)	0.908 (0.04)	0.918 (0.04)	0.923 (0.03)

Note: Bold value represents the best dice scores among different methods.

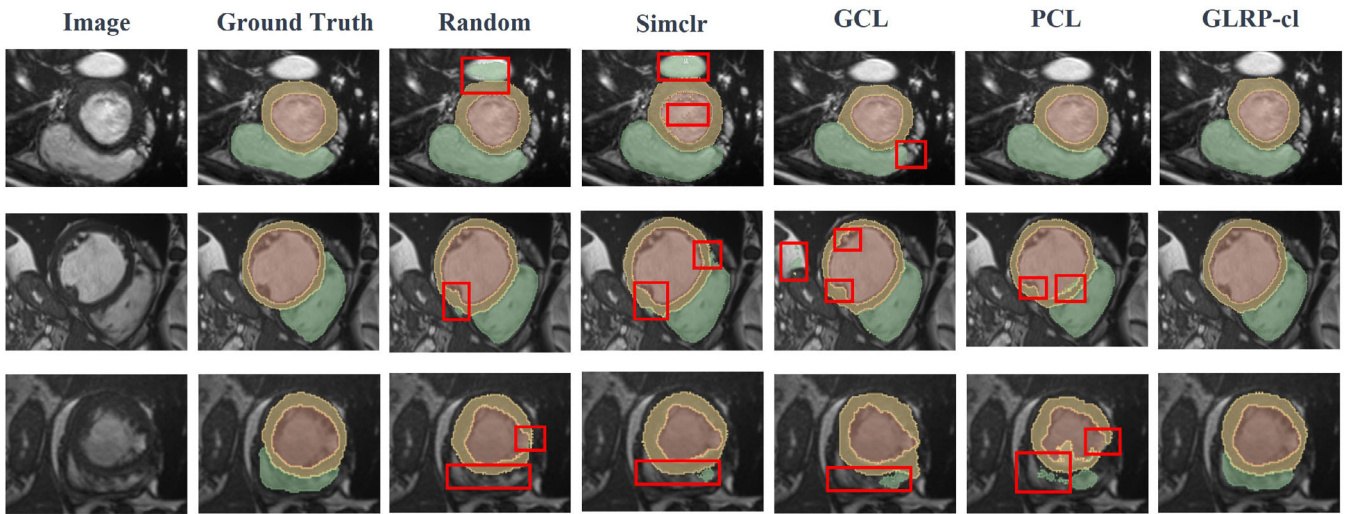


FIGURE 5 Comparison of segmentation results for the ACDC dataset.

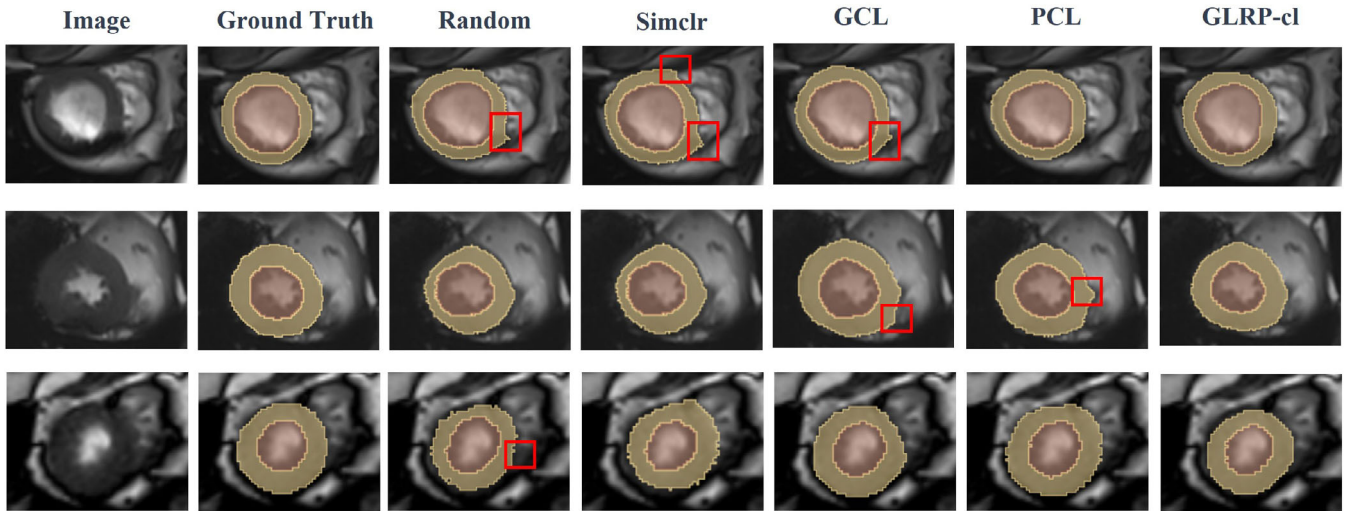


FIGURE 6 Comparison of segmentation results for the private dataset.

4.4.2 | Comparative experiments on the definition of similarity in global contrastive strategy

In short-axis cardiac MRI images, the slice plane has a low spatial resolution, and the similarity of images decreases sharply with increasing position differences. Therefore, in 3.3.2, we proposed a strategy to define this similarity based on relative position. In order to demonstrate the effectiveness of this strategy, a comparison experiment called GLRP-cl* was designed in this section, which uses the linear position difference shown in the following equation to define the similarity between slices.

$$d_{ij} = x_i - x_j \quad (10)$$

x_i and x_j are the positional labels for slices i and j . In this experiment, we kept $k = 10$. At the same time, the results of the PCL method were used as a baseline for comparison. The experimental results of the ACDC dataset and the private dataset are shown in Tables 3 and 4, respectively.

The performance of GLRP-cl is superior to that of GLRP-cl*, which demonstrates the effectiveness of the proposed similarity based on relative position. Therefore, we can conclude that this method of defining the relative positive and negative samples based on the actual medical image context can facilitate the pre-trained model to learn high-quality sample features.

TABLE 3 Metrics of the segmentation results for the ACDC dataset.

Method	Dice				Hausdorff distance (mm)			
	LV	Myo	RV	Mean (STD)	LV	Myo	RV	Mean (STD)
PCL	0.8813	0.7817	0.7389	0.801 (0.07)	9.07	20.52	22.10	17.23 (7.1)
GLRP-cl*	0.8924	0.8208	0.7724	0.829 (0.06)	9.36	9.91	18.62	12.63 (5.2)
GLRP-cl	0.9227	0.8280	0.8026	0.851 (0.06)	7.31	11.39	16.15	11.62 (4.4)

Note: Bold value represents the best values among different methods.

TABLE 4 Metrics of the segmentation results for the private dataset.

Method	Dice				Hausdorff distance (mm)			
	LV	Myo	RV	Mean (STD)	LV	Myo	RV	Mean (STD)
PCL	0.9321	0.8178	0.8058	0.852 (0.07)	4.90	6.48	10.23	7.20 (2.7)
GLRP-cl*	0.9132	0.8123	0.8224	0.849 (0.06)	12.26	5.08	12.26	9.87 (4.1)
GLRP-cl	0.9339	0.8302	0.8443	0.869 (0.06)	3.75	4.60	7.20	5.18 (1.8)

Note: Bold value represents the best values among different methods.

TABLE 5 Metrics of the segmentation results for the ACDC dataset.

Method	Dice				Hausdorff distance (mm)			
	LV	Myo	RV	Mean (STD)	LV	Myo	RV	Mean (STD)
PCL	0.8813	0.7817	0.7389	0.801 (0.07)	9.07	20.52	22.10	17.23 (7.1)
GRP-cl	0.9015	0.7857	0.7496	0.812 (0.08)	8.59	11.90	17.52	12.67 (4.5)
LRP-cl	0.8832	0.8089	0.7612	0.818 (0.06)	8.37	12.14	24.40	14.97 (8.4)
GLRP-cl	0.9227	0.8280	0.8026	0.851 (0.06)	7.31	11.39	16.15	11.62 (4.4)

Note: Bold value represents the best values among different methods.

TABLE 6 Metrics of the segmentation results for the private dataset.

Method	Dice				Hausdorff distance (mm)			
	LV	Myo	RV	Mean (STD)	LV	Myo	RV	Mean (STD)
PCL	0.9321	0.8178	0.8058	0.852 (0.07)	4.90	6.48	10.23	7.20 (2.7)
GRP-cl	0.9193	0.8208	0.8677	0.869 (0.05)	3.59	4.32	8.38	5.43 (2.6)
LRP-cl	0.9255	0.8184	0.8344	0.859 (0.06)	3.73	4.46	9.31	5.83 (3.0)
GLRP-cl	0.9339	0.8302	0.8443	0.869 (0.06)	3.75	4.60	7.20	5.18 (1.8)

Note: Bold value represents the best values among different methods.

4.4.3 | Ablation experiments on GCL and LCL

In this section, experiments will be performed on two aspects. First, an ablation experiment is performed. The results of PCL are used as the baseline results, and the

effectiveness of each module is verified by comparing the results of GRP-cl, LRP-cl, and GLRP-cl. In this section, $k = 10$ is used. The results of the ACDC dataset and the private dataset are shown in Tables 5 and 6, respectively.

Overall, the GLRP-cl method achieves the best segmentation results, indicating that the proposed strategy

TABLE 7 Metrics of the segmentation results for the ACDC dataset.

Method	Dice				Hausdorff distance (mm)			
	LV	Myo	RV	Mean (STD)	LV	Myo	RV	Mean (STD)
GRP-cl	0.9015	0.7857	0.7496	0.812 (0.08)	8.59	11.90	17.52	12.67 (4.5)
LRP-cl	0.8832	0.8089	0.7612	0.818 (0.06)	8.37	12.14	24.40	14.97 (8.4)
GLRP-cl	0.9227	0.828	0.8026	0.851 (0.06)	7.31	11.39	16.15	11.62 (4.4)
GLRP-cl-crf	0.9274	0.8472	0.8534	0.876 (0.04)	21.33	19.60	38.27	26.40 (10.3)

Note: Bold value represents the best values among different methods.

TABLE 8 Metrics of the segmentation results for the private dataset.

Method	Dice				Hausdorff distance (mm)			
	LV	Myo	RV	Mean (STD)	LV	Myo	RV	Mean (STD)
GRP-cl	0.9193	0.8208	0.8677	0.869 (0.05)	3.59	4.32	8.38	5.43 (2.6)
LRP-cl	0.9255	0.8184	0.8344	0.859 (0.06)	3.73	4.46	9.31	5.83 (3.0)
GLRP-cl	0.9339	0.8302	0.8443	0.869 (0.06)	3.75	4.60	7.20	5.18 (1.8)
GLRP-cl-crf	0.9483	0.8556	0.8940	0.899 (0.05)	5.01	6.98	9.15	7.05 (2.1)

Note: Bold value represents the best values among different methods.

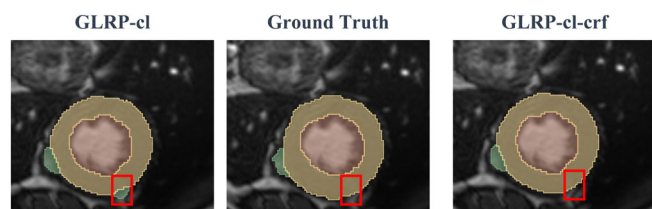


FIGURE 7 Comparison of segmentation results for the private dataset.

enables the model's encoder to learn sample-specific structural information and higher-order features, thus providing strong support for the segmentation task.

4.4.4 | Comparative experiments on CRF

In this section, a comparative experiment is performed. The results of GLRP-cl are used as a baseline, and the effectiveness of fully connected CRF is verified by comparing the baseline with the results of post-processing GLRP-cl by CRF. The experimental results of the ACDC dataset and the private dataset are shown in Tables 7 and 8, respectively.

As shown in Figure 7, the CRF module was able to eliminate the isolated mis-segmentations generated by the GLRP-cl method. Tables 7 and 8 show that the post-processing of GLRP-cl with CRF achieves better results

than GLRP-cl, GRP-cl, and LRP-cl, proving that the CRF module can improve the segmentation accuracy.

5 | DISCUSSION AND CONCLUSION

In this study, we propose a global and local contrastive learning method based on relative position to guide the segmentation network to learn the characteristics of the training samples through a clustering task based on position information in the pre-training stage, and then complete the segmentation task based on a small number of training samples in the fine-tuning stage. In the post-processing stage, we further optimize the segmentation results using a fully connected CRF model to reduce the mis-segmented regions and achieve better segmentation results. Our proposed method is evaluated on both public and private datasets, and the results show that the method can improve the segmentation performance with limited labeled data.

It's important to acknowledge the challenges we encountered in implementing our experiments. One of the main challenges is the high computational cost of the pre-training network, which requires a large amount of training data and computational resources. Each pre-training takes 1–2 days. To overcome this challenge, we employ several techniques to reduce the training time and memory consumption, such as data augmentation,

batch normalization, and early stopping. Another challenge is the time-consuming and labor-intensive selection and labeling of the private dataset. The medical experts at Beijing Anzhen Hospital have spent a large amount of time on this task. Besides, there are some limitations to our work: we only perform the experiments on short-axis cardiac MRI images and have not applied them to other medical imaging modalities. In addition, the definition of relative positive and negative samples relies solely on the relative positions of image slices and feature map patches within 3D medical images, omitting other important image features. Furthermore, while the study effectively addresses training strategies for scenarios with limited annotated data, it does not extensively explore methods to improve segmentation accuracy, and the pre-training stage incurs significant time costs. Future work should extend the applicability of the method to different imaging modalities, incorporate additional image features, refine segmentation accuracy, explore parallelization for efficient model training, and focus on optimizing network architectures to improve efficiency and reduce training time.

ACKNOWLEDGMENTS

This work was supported in part by the ERC IMI (101005122), the H2020 (952172), the Medical Research Council (MRC) (MC/PC/21013), the Royal Society (IEC \NSFC\211235), the NVIDIA Academic Hardware Grant Program, the SABER project supported by Boehringer Ingelheim Ltd, and the UKRI Future Leaders Fellowship (MR/V023799/1).

CONFLICT OF INTEREST STATEMENT

The authors declare that there are no conflicts of interest.

DATA AVAILABILITY STATEMENT

Research data are not shared.

ORCID

Xin Zhao  <https://orcid.org/0000-0002-1972-3669>

Tongming Wang  <https://orcid.org/0009-0000-8029-2514>

Jingsong Chen  <https://orcid.org/0009-0001-7390-8330>

Bingrun Jiang  <https://orcid.org/0009-0008-1227-7951>

Haotian Li  <https://orcid.org/0009-0007-9676-7220>

Nan Zhang  <https://orcid.org/0000-0002-8309-4262>

Guang Yang  <https://orcid.org/0000-0001-7344-7733>

Senchun Chai  <https://orcid.org/0000-0003-1910-1795>

REFERENCES

- Kaus MR, von Berg J, Weese J, Niessen W, Pekar V. Automated segmentation of the left ventricle in cardiac MRI. *Med Image Anal.* 2004;8(3):245-254. doi:10.1016/j.media.2004.06.015
- Khened M, Kollerathu VA, Krishnamurthi G. Fully convolutional multi-scale residual DenseNets for cardiac segmentation and automated cardiac diagnosis using ensemble of classifiers. *Med Image Anal.* 2019;51:21-45. doi:10.1016/j.media.2018.10.004
- Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* IEEE; 2015:3431-3440.
- Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF, eds. *Lecture Notes in Computer Science Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015.* Vol 9351. Springer International Publishing; 2015:234-241. doi:10.1007/978-3-319-24574-4_28
- Isensee F, Petersen J, Klein A, et al. Nnu-net: self-adapting framework for u-net-based medical image segmentation [J]. *arXiv Preprint arXiv:1809.10486.* 2018.
- Sun X, Garg P, Plein S, van der Geest RJ. SAUN: stack attention U-net for left ventricle segmentation from cardiac cine magnetic resonance imaging. *Med Phys.* 2021;48(4):1750-1763. doi:10.1002/mp.14752
- He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* IEEE; 2016:770-778.
- Huang G, Liu Z, van der Maaten L, et al. Densely connected convolutional networks[C]. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* IEEE; 2017:4700-4708.
- Oktay O, Schlemper J, Folgoc LL, et al. Attention u-net: learning where to look for the pancreas[J]. *arXiv Preprint arXiv:1804.03999.* 2018.
- Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. *Adv Neural Inf Proces Syst.* 2017;30:5998-6008.
- Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: transformers for image recognition at scale [J]. *arXiv Preprint arXiv:2010.11929.* 2020.
- Chen J, Lu Y, Yu Q, et al. Transunet: transformers make strong encoders for medical image segmentation[J]. *arXiv Preprint arXiv:2102.04306.* 2021.
- Cao H, Wang Y, Chen J, et al. Swin-Unet: Unet-Like Pure Transformer for Medical Image Segmentation[C]. *Computer Vision-ECCV 2022 Workshops.* 2023 205-218.
- Liu X, Zhang F, Hou Z, et al. Self-supervised learning: generative or contrastive. *IEEE Trans Knowl Data Eng.* 2021;35(1):857-876.
- Sun L, Zhang M, Wang B, Tiwari P. Few-shot class-incremental learning for medical time series classification. *IEEE J Biomed Health Inform.* 2023;1-11. doi:10.1109/JBHI.2023.3247861
- Wang W, Chen Z, Yuan X. Simple low-light image enhancement based on weber-Fechner law in logarithmic space[J]. *Signal Process Image Commun.* 2022;106:116742.
- Liu M, Zhang X, Yang B, et al. Three-dimensional modeling of heart soft tissue motion[J]. *Appl Sci.* 2023;13(4):2493.
- Lu S, Liu S, Hou P, et al. Soft tissue feature tracking based on DeepMatching network[J]. *CMES-Comput Model Eng Sci.* 2023; 136(1):363-379.
- Liu Y, Tian J, Hu R, et al. Improved feature point pair purification algorithm based on SIFT during endoscope image stitching[J]. *Front Neurorobot.* 2022;16:840594.

20. Qadri F, Lin H, Shen L, et al. CT-based automatic spine segmentation using patch-based deep learning[J]. *Int J Intell Syst*. 2023;2023:1-14.
21. Hirra I, Ahmad M, Hussain A, et al. Breast cancer classification from histopathological images using patch-based deep learning modeling[J]. *IEEE Access*. 2021;9:24273-24287.
22. Jaiswal A, Babu AR, Zadeh MZ, Banerjee D, Makedon F. A survey on contrastive self-supervised learning. *Dent Tech*. 2020; 9(1):2. doi:10.3390/technologies9010002
23. Zeng D, Wu Y, Hu X, et al. Positional contrastive learning for volumetric medical image segmentation. *Medical Image Computing and Computer Assisted Intervention-MICCAI 2021: 24th International Conference*, September 27–October 1, 2021, Proceedings, Part II 24. Springer; 2021:221-230.
24. Chaitanya K, Erdil E, Karani N, Konukoglu E. Contrastive learning of global and local features for medical image segmentation with limited annotations. *Adv Neural Inf Proces Syst*. 2020;33:12546-12558.
25. Wang W, Zhou T, Yu F, Dai J, Konukoglu E, van Gool L. Exploring cross-image pixel contrast for semantic segmentation. *Proceedings of the IEEE/CVF International Conference on Computer Vision*. IEEE; 2021:7303-7313.
26. You C, Zhou Y, Zhao R, Staib L, Duncan JS. Simcvd: simple contrastive voxel-wise representation distillation for semi-supervised medical image segmentation. *IEEE Trans Med Imaging*. 2022;41(9):2228-2237. doi:10.1109/TMI.2022.3161829
27. Liu Z, Li Z, Hu Z, et al. Contrastive and selective hidden embeddings for medical image segmentation. *IEEE Trans Med Imaging*. 2022;41(11):3398-3410. doi:10.1109/TMI.2022.3186677
28. Chaitanya K, Erdil E, Karani N, Konukoglu E. Local contrastive loss with pseudo-label based self-training for semi-supervised medical image segmentation[J]. *Med Image Anal*. 2023;102792:102792.
29. Xiang J, Li Z, Wang W, Xia Q, Zhang S. "Self-ensembling contrastive learning for semi-supervised medical image segmentation," ArXiv Prepr ArXiv210512924. 2021.
30. Krhenbühl P, Koltun V. *Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials*. Curran Associates Inc; 2012.
31. Khosla P, Teterwak P, Wang C, et al. Supervised contrastive learning. *Adv Neural Inf Proces Syst*. 2020;33:18661-18673.
32. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. *IEEE Trans Image Process*. 2004;13(4):600-612. doi:10.1109/TIP.2003.819861
33. Bernard O, Lalande A, Zotti C, et al. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE Trans Med Imaging*. 2018;37(11):2514-2525. doi:10.1109/TMI.2018.2837502
34. Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. *International Conference on Machine Learning*. PMLR; 2020:1597-1607.

How to cite this article: Zhao X, Wang T, Chen J, et al. GLRP: Global and local contrastive learning based on relative position for medical image segmentation on cardiac MRI. *Int J Imaging Syst Technol*. 2023;1-14. doi:10.1002/ima.22992