J. Shanghai Jiao Tong Univ. (Sci.), 2023
https://doi.org/10.1007/s12204-023-2614-y

Positional Information is a Strong Supervision for Volumetric Medical Image Segmentation

ZHAO Yinjie[1] (赵寅杰), *HOU Runping*[1] (侯润萍), *ZENG Wanqin*[2] (曾琬琴), *QIN Yulei*[1] (秦玉磊), *SHEN Tianle*[2] (沈天乐), *XU Zhiyong*[2] (徐志勇), *FU Xiaolong*[2*] (傅小龙), *SHEN Hongbin*[1*] (沈红斌)
(1. School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China;
2. Shanghai Chest Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai 200030, China)

Abstract: Medical image segmentation is a crucial preliminary step for a number of downstream diagnosis tasks. As deep convolutional neural networks successfully promote the development of computer vision, it is possible to make medical image segmentation a semi-automatic procedure by applying deep convolutional neural networks to finding the contours of regions of interest that are then revised by radiologists. However, supervised learning necessitates large annotated data, which are difficult to acquire especially for medical images. Self-supervised learning is able to take advantage of unlabeled data and provide good initialization to be finetuned for downstream tasks with limited annotations. Considering that most self-supervised learning especially contrastive learning methods are tailored to natural image classification and entail expensive GPU resources, we propose a novel and simple pretext-based self-supervised learning method that exploits the value of positional information in volumetric medical images. Specifically, we regard spatial coordinates as pseudo labels and pretrain the model by predicting positions of randomly sampled 2D slices in volumetric medical images. Experiments on four semantic segmentation datasets demonstrate the superiority of our method over other self-supervised learning methods in both semi-supervised learning and transfer learning settings. Codes are available at https://github.com/alienzyj/PPos.
Key words: self-supervised learning, medical image analysis, semantic segmentation
CLC number: TP 39, R 445 **Document code:** A

0 Introduction

In medical image analysis, semantic segmentation is a critical procedure, as it is a preceding step for multiple downstream tasks. Over recent years, as deep convolutional neural networks (DCNNs) gained in popularity in all sorts of computer vision tasks, hundreds of models have been proposed to push back the frontiers of semantic segmentation[1]. Therefore, in medical image analysis, there is a growing trend towards the utilization of these deep learning models to find prospective regions of interest (ROIs) that will be inspected by radiologists subsequently[2]. Nevertheless, to attain promising results, supervised learning requires large labelled training data, which are extremely difficult to collect and annotate, especially for medical images.

Considering that there are lots of unlabelled images during routine clinical diagnosis, self-supervised learning is highly desirable to take advantage of these data. A number of self-supervised learning methods have been developed in recent years[3], among which contrastive learning is the most prevalent and predominant[4-6]. Although the power of contrastive learning has been demonstrated mainly in natural image classification, these methods are not easy to directly transfer to medical image segmentation for the following reasons. First, most contrastive learning methods define positive/negative samples by instance discrimination[7], which may incur inferior performance for medical image segmentation because regarding each medical image as an independent class contradicts the clinical settings and fails to meet the demand of voxel-level segmentation tasks. Second, contrastive learning enormously relies on heavy data augmentations[4], whereas several of them are not feasible to apply to medical images. For example, excessive deformation or improper density transformation may induce undesirable or even nonexistent anatomical patterns. Thus, it is highly desirable to propose a method that is fit for the special characteristics of medical images and exploits clinical prior knowledge.

In clinical scenarios, it is noticeable that radiologists tend to refer to the position of slices in 3D medical

Received: 2022-08-08 **Accepted:** 2022-11-28
Foundation item: the Major Research Plan of the National Natural Science Foundation of China (No. 92059206)
*__E-mail:__ xlfu1964@hotmail.com; hbshen@sjtu.edu.cn

images when deciding whether the objects in the foreground are lesions or not. For example, brachiocephalic trunk, left common carotid artery, and left subclavian artery should be observed at the height of the fifth thoracic, despite some visual discrepancies in different patients. If there are some anomalies that are not supposed to appear at certain position such as solid ground glass objects in the thoracic cavity, they are very likely to become lesions. Motivated by this human expertise, we aim to formulate a pretext task suitable for medical image analysis by taking above domain knowledge into account.

In this work, we develop a novel pretext-based self-supervised learning method whose pseudo labels are generated from spatial coordinates. Our inspiration comes from the prior knowledge that the relationship between the semantics of image contents and corresponding positional information is fairly close in medical images. Specifically, medical images are mainly composed of anatomical structures such as tissues and organs that are consistent in their locations across different patients. Therefore, this relationship can be exploited in self-supervised learning. For example, several previous works[8-9] utilized absolute or relative positions to define the positive/negative samples in contrastive learning and achieved promising results in medical image segmentation. To prove that positional information can be regarded as strong supervision for DCNNs, we propose taking advantage of positional coordinates of 2D slices extracted from volumetric medical images at self-supervised learning stage. This method incorporates positional prior knowledge and improves the performance on organ and lesion segmentation tasks. The pretext task is quite simple and straightforward to implement, and it has the potential to generalize to a wide variety of medical image modalities and multiple clinical diagnosis tasks.

Our contributions are mainly twofold:

(1) We propose a novel and simple self-supervised learning method called predicting positions (PPos) by using positional coordinates as pseudo labels. It avoids expensive computational overheads such as large GPU memory and long training duration that most contrastive learning methods necessitate.

(2) We demonstrate the effectiveness and competence of our method compared with other self-supervised learning methods through experiments on two in-house lung nodule segmentation datasets and two public cardiac substructure segmentation datasets.

1 Related Work

1.1 Medical Image Segmentation

Medical image segmentation has been researched for years[1]. The most frequently used and studied DCNN in medical image segmentation is U-Net[10]. It represents a paradigm of using the encoder-decoder architecture with skip connection for 2D medical image segmentation. V-Net[11] and 3D U-Net[12] transferred the network to 3D settings and demonstrated the capability to handle 3D inputs. From then on, there have been dozens of papers trying to modify the network architecture in order to realize the potential of U-Net[13-14]. Additionally, nnU-Net[15] unified the general preprocessing and training procedure of U-Net and attained state-of-the-art results on multiple datasets.

1.2 Self-Supervised Learning

Self-supervised learning has been thoroughly investigated over recent years[3]. Since there are much more unlabelled data in real-world settings, self-supervised learning is a popular pretraining step to utilize these unlabelled data by providing proper initialization for downstream tasks. There are mainly three types of self-supervised learning[3]. In early research, pretext-based methods were first proposed such as solving jigsaw puzzles[16] and context prediction[17]. Another type is generative methods such as image colorization[18] and pixel reconstruction[19]. However the most popular kind of self-supervised learning is contrastive learning, which has dominated multiple vision tasks in the last few years[4−6]. Most contrastive learning methods utilize instance discrimination[7] and InfoNCE is the most common loss function[5,20]. Contrastive learning usually entails a large number of negative samples to avoid trivial solution[4] and thus is quite expensive to train a model. MoCo[5] introduced a query queue to store the embedding of negative samples, which made the large batch size dispensable during pretraining. In addition, BYOL[6] eliminated negative samples in the loss function and SimSiam[21] further simplified the paradigm of contrastive learning.

Due to the difference between natural and medical images, several self-supervised learning methods have been specially designed for medical image analysis. Genesis[22-23] enforced the model to reconstruct voxels painted or distorted by preprocessing, while Refs. [24-25] used Rubik's cube recovery as their pretext task. TranferVW[26] clustered unlabelled chest X-rays by latent features of an autoencoder and randomly selected several patches at the same location among these clustered samples as their visual words. An encoder-decoder model was subsequently pretrained by visual word classification and pixel reconstruction tasks simultaneously. Moreover, GCL[8] and PCL[9] were two contrastive methods that defined positive/negative samples by absolute or relative positions.

2 Method

2.1 Presupposition

Our method is based on a distinctive feature peculiar to medical images. For medical images aiming at

specific anomalies in one modality such as CT scans for lung nodules, the anatomical structures are similar across different patients in general[8-9]. Although statures, genders, and ages have some effect on the visual patterns of anatomical structures, our method is based on the correlation between positional and anatomical information instead of identical shapes of organs in different patients. Specifically, if we align two volumes and extract one slice of each volume at the same position, consistent anatomical structures could be observed between the slices since they contain the same region that covers an identical collection of tissues and organs. It indicates that anatomical information is correlated with positional information in medical images, which can be utilized as a supervision signal in self-supervised learning. We presume that the correlation between specific anatomical structure and its corresponding position is supposed to benefit downstream tasks such as finding contours of target structures in a specific region.

2.2 Pretext Task

In this work, we propose a novel and simple pretext task that directly uses normalized z-axis coordinates of 2D slices extracted from volumetric medical images as pseudo labels. Since CT scans covering different regions result in inconsistency of coordinate systems, the alignment is a critical step in the validity of the pseudo labels. During preprocessing, all the medical images are resampled into the same voxel spacing and the body region is cropped. This body-centred crop has two main purposes. One is to eliminate the superfluous part outside the body of patient to reduce the input size and speed up the process. The other is to make sure the consistency of anatomical structures with corresponding normalized coordinates across different medical images, which is an essential premise of our method. The above alignment may not be highly precise, but it performs well in our experiments and ensures the simplicity of preprocessing. Subsequently, we randomly extract 2D slices from the above cropped volumetric medical images, along with their normalized z-axis coordinates as shown in Fig. 1. At pretraining stage, these 2D slices are propagated to the model, while the corresponding coordinates are used as supervision to train the encoder of U-Net followed by a predictor. After pretraining, the encoder is finetuned with a decoder for downstream tasks whereas the predictor is discarded.

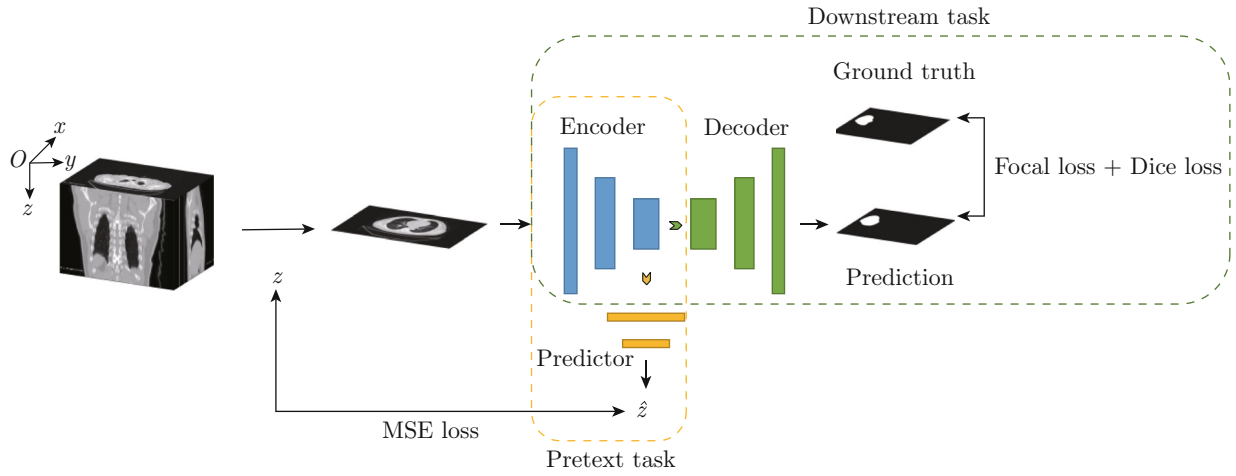


Fig. 1 Sketch of the proposed PPos framework

It is worth noting that several previous works such as TransVW[26], GCL[8], PCL[9], Jigsaw[16], UBR[27], and MVP-Net[28] also utilized positional information. The main difference between our method and these methods is that we explicitly take advantage of positional information by regarding spatial coordinates of slices as pseudo labels, whereas previous works exploited positional information in an indirect way. Specifically, TransVW randomly selected patches at the same coordinate across patients clustered by latent features of an autoencoder as visual words. Instances of each visual word exhibited a high degree of similarity in anatomical patterns at the same position. Therefore, positional information was implicitly embodied in the visual word classification task. In terms of GCL and PCL, these two methods defined positive/negative samples by absolute or relative positions within contrastive learning framework, which meant that positional information was represented in the contrastive loss function rather than directly used as supervision signals. In the jigsaw task, the input data were a collection of identical square tiles from an image and the labels were permutations of these tiles taken from a predefined set[16]. The model learned the positional information by finding the correlation of each separate puzzle tile to a precise object part during jigsaw puzzle reassembly. As for UBR, it

did not align the CT scans of different patients, which meant the slice indices could not be directly used as z-axis coordinates or as labels. Instead, it enforced the deeply learned regressor to obey the spatial superior-inferior ordering as a hard constraint, and ensured that the predicted slice scores were approximately proportional to the spatial distance between slice indices[27]. The above two methods enforced the model to learn positional information by indirect pretext tasks instead of normalized coordinates. In terms of MVP-Net, it proposed to learn position-aware appearance features by introducing a position prediction task during training, in order to embed the position information onto the appearance features. Specifically, entangled position and appearance features were learned through the multi-task design in the MVP-Net that jointly predicted the position and the detection bounding box[28]. However our method exploits positional information at the pretraining stage before finetuning for downstream tasks, which means it decouples the procedure to learn positional information from solving semantic segmentation problems and it can take advantage of unlabelled data.

2.3 Learning Procedure

As shown in Fig. 1, our method follows the general self-supervised learning process. At pretraining stage, we first extract the encoder of U-Net and concatenate a multi-layer perceptron to it as a predictor. Then, we randomly sample 2D slices from volumetric medical images followed by data augmentations as inputs. After the 2D slices are projected into the embedding space by the encoder, these embeddings are subsequently processed by the predictor to predict coordinates. Mean square error (MSE) loss is calculated on the groundtruth and predicted coordinates to update the parameters of the model:

$$\mathrm{MSELoss}(z, \hat{z}) = \frac{1}{N}\sum_{i=1}^{N}(z_i - \hat{z}_i)^2, \tag{1}$$

where z and $\hat{z}$ respectively refer to the groundtruth and the predicted z-axis coordinates in one mini-batch, and N is the batch size. Once the pretraining is finished, the encoder should provide appropriate initialization to be finetuned together with the decoder for downstream tasks, whereas the predictor is discarded. During finetuning, we suppose that the model has learned the correlation between anatomical and positional information, which is not restricted to any specific coordinate system. Thus, it is not necessary to align samples in downstream datasets.

3 Experiments

3.1 Setup

3.1.1 *Datasets*

Experiments are implemented on two in-house datasets and two public datasets. Two in-house datasets are composed of CT scans with lung nodules. The first dataset, LargeThick, consists of 989 CT scans for training and 119 for testing with slice thickness around 5 mm. Lung nodules in this dataset are diverse and comprise hard samples such as nodules overlapping adjacent tissues. The other dataset, SmallThin, consists of 172 CT scans for training and 21 for testing with slice thickness under 1.5 mm. In this dataset, lung nodules are more likely to be self-isolated. In both datasets, each CT scan only contains one nodule.

Two public datasets consist of CT scans covering cardiac region. The first one is CHD[29] dataset, which consists of 91 3D cardiac CT scans and covers 14 types of congenital heart disease including seven substructures: left ventricle, right ventricle, left atrium, right atrium, myocardium, aorta, and pulmonary artery. The second dataset is MMWHS[30-31] composed of 20 CT scans and seven substructures as in CHD.

At self-supervised learning stage, all the samples in the training set are viewed as unlabelled data to pretrain the model, whereas for downstream tasks, we just use a limited number of training data with annotations to finetune the model.

3.1.2 *Networks*

For in-house datasets, we use three models: a 2D U-Net, a 3D U-Net and a classifier. The 2D U-Net and 3D U-Net are variants of U2-Net[32], while the classifier is based on ResNet[33]. In addition, to improve the accuracy of ROI masks, we utilize malignant/benign nodule classification as an auxiliary task when training the 3D U-Net. These three models are trained individually and then cascaded during testing. We apply self-supervised learning methods to 2D U-Net in our experiments whereas the 3D U-Net and classifier are trained from scratch. For public datasets, we apply a simple 2D U-Net to finding the contours of cardiac substructures in each 2D slice extracted from 3D CT scans.

3.1.3 *Preprocessing*

For in-house datasets, the raw file format of the CT scans is DICOM, which contains not only the voxel arrays but also their positions in the world coordinate system. So the slices are reordered according to their positions when we build our datasets, which ensures that the direction of CT scans is consistent among different patients. In order to generalize U-Net to a wide range of CT scans, we unify the voxel spacing of all samples in both datasets by resampling into {0.7 mm, 0.7 mm, 1.5 mm} for $\{x, y, z\}$ axes respectively. After resampling, we use digital image processing algorithm to find the coarse mask of the lung and crop this region to ensure that the anatomical contents are consistent across samples. All the voxel values of CT scans in our in-house datasets are cropped by lung window $[-1\,024, 400]$ and then normalized to $[0, 1]$. For public datasets, voxel spacing is resampled into {1.0 mm,

1.0 mm, 1.0 mm} and voxel values are normalized by the mean and standard of each image.

3.1.4 *Training Details*

For in-house datasets, models are optimized by stochastic gradient descent (SGD) with initial learning rate 0.01 and cosine scheduler during pretraining. Data augmentations are applied including elastic transformation, random rotation, random crop and random flip. We pretrain models by PPos for 800 epochs and all the other methods for 300 epochs on one NVIDIA GeForce GTX 2080Ti GPU due to their different training speeds and convergence points. For downstream tasks, we finetune these models by SGD with initial learning rate 0.01 for 100 epochs. Dice loss[11] and focal loss[34] are used during finetuning:

$$\text{DiceLoss}(p,y)=1-\frac{\sum_i p_i y_i}{\sum_i p_i^2+\sum_i y_i^2-\sum_i p_i^2 y_i^2}, \quad (2)$$

$$\text{FocalLoss}(p,y)=-\sum_i \hat{\alpha}(1-\hat{p}_i)^{\gamma}\log \hat{p}_i, \quad (3)$$

$$\hat{p}_i=\begin{cases} p_i, & y_i=1 \\ 1-p_i, & y_i=0 \end{cases}, \quad (4)$$

$$\hat{\alpha}_i=\begin{cases} \alpha, & y_i=1 \\ 1-\alpha, & y_i=0 \end{cases}, \quad (5)$$

where y and p refer to the groundtruth and predicted ROI masks respectively, in which $\hat{y}$ and $\hat{p}$ are labels for each voxel. Additionally, α is a weighting factor set by inverse class frequency, whereas γ is a tunable modulating factor that reduces the loss contribution from easy samples and extends the range in which an sample receives low loss[34].

For public datasets, at pretraining stage models are optimized by SGD with initial learning rate 0.1 and batch size 32 for 200 epochs on two NVIDIA GeForce GTX 2080Ti GPUs. Cosine scheduler is used to improve the results. Data augmentations are applied including elastic transformation, random rotation, and random flip. For downstream tasks, we finetune these models by SGD with initial learning rate 5×10^{-5} for 100 epochs. Cross entropy is used as loss function during finetuning.

3.1.5 *Inference*

For in-house datasets, the input of our pipeline during inference is the whole CT scan. After preprocessing, we split the scan into slices and apply 2D U-Net to these slices to find prospective nodules. Subsequently, these suspected nodules are processed by the classifier to determine the most probable one and discard all the others in order to reduce false positives since each CT scan in both datasets only has one target nodule. Finally, a refined mask of the reserved nodule is drawn by 3D U-Net, which is the definitive output of our pipeline. For public datasets, we split each volumetric CT scan into 2D slices and process each slice individually. After preprocessing, these slices are propagated into the 2D U-Net to generate masks for cardiac substructures.

3.2 Results

In the following experiments, we compare our method PPos with training from scratch Random, Genesis[22-23], Rotation[35], MoCo[5], PIRL[36], SimCLR[4], GCL[8], and PCL[9]. All the models except Random are pretrained on the entire training set of the pretraining dataset without using annotations and finetuned by different numbers of training samples with annotations on the downstream dataset. The downstream dataset is either the same as pretraining dataset or another transferring dataset, referred to as semi-supervised learning and transfer learning respectively.

3.2.1 *Semi-Supervised Learning*

As shown in Table 1, when pretraining and finetuning on LargeThick dataset, our method is competitive with other methods on different percentages of training set. Moreover, PPos demonstrates its superiority to contrastive methods GCL and PCL that also utilize positional information. We speculate that the assumption in GCL and PCL that positive and negative samples can be simply defined by positional difference along z-axis may not hold firm here, since the similarity

Table 1 Comparison of different self-supervised methods pretraining and finetuning on LargeThick dataset

Method	10%			50%			100%		
	Dice	P	R	Dice	P	R	Dice	P	R
Random	0.509 9	0.503 6	**0.607 2**	0.620 6	0.641 9	**0.660 5**	0.663 3	0.670 2	0.702 4
Genesis[22-23]	0.541 4	**0.615 5**	0.573 8	0.634 9	0.672 3	0.657 1	0.677 3	0.673 7	**0.725 3**
GCL[8]	0.523 9	0.577 0	0.551 8	0.623 1	0.657 2	0.646 5	0.675 6	0.682 2	0.713 9
PCL[9]	0.540 0	0.585 6	0.561 8	0.628 8	0.658 4	0.649 6	0.674 2	0.680 7	0.708 8
PPos	**0.542 9**	0.602 4	0.571 9	**0.636 3**	**0.673 4**	0.656 0	**0.681 2**	**0.686 3**	0.718 9

Note: 10%, 50%, and 100% refer to the percentages of training samples when the model is finetuned; Dice, P, and R refer to dice coefficient, average voxel precision, and average voxel recall respectively.

between slices may not be exactly consistent with the manually selected partition in GCL or threshold in PCL. PPos provides an alternative to exploit positional information more flexibly and therefore avoids such confusion.

Results on CHD dataset are shown in Table 2. Although the performance is inferior to PCL and GCL when $M = 2$, PPos still outperforms other methods when more training samples are provided. It is notable that another pretext-based method Rotation does not perform well on both datasets, since it is designed for natural images and does not take medical domain knowledge into account compared with ours. Besides, it can be observed that MoCo, PIRL, and SimCLR do not perform very well and even get lower dice coefficients than Random when finetuning on $M \geqslant 10$. We attribute it to the utilization of instance discrimination in SimCLR, which is not suitable for medical image segmentation and even has an adverse effect on the performance. By contrast, GCL, PCL, and PPos that exploit positional information demonstrate an overall improvement over training from scratch.

3.2.2 *Transfer Learning*

Results for models pretrained on LargeThick and finetuned on SmallThin are shown in Table 3. We find that when transferring to small dataset, PPos can attain significant gain compared with other methods. Specifically, the model pretrained by PPos and finetuned on 50% of training samples can achieve higher dice coefficient than models pretrained by other methods and finetuned on the whole training set. It reveals the potential of PPos in transfer learning and indicates that positional information embodies strong supervision to pretrain deep learning models.

Table 2 Comparison of different methods pretraining and finetuning on CHD dataset

Method	$M = 2$	$M = 10$	$M = 25$	$M = 50$	$M = 72$
Random	0.147 ± 0.06	0.549 ± 0.04	0.690 ± 0.04	0.725 ± 0.03	0.757 ± 0.03
Rotation[35]	0.166 ± 0.07	0.507 ± 0.06	0.621 ± 0.05	0.702 ± 0.04	0.740 ± 0.04
MoCo[5]	0.181 ± 0.07	0.516 ± 0.05	0.630 ± 0.06	0.711 ± 0.04	0.745 ± 0.04
PIRL[36]	0.191 ± 0.06	0.535 ± 0.04	0.668 ± 0.05	0.724 ± 0.04	0.752 ± 0.03
SimCLR[4]	0.180 ± 0.07	0.522 ± 0.04	0.640 ± 0.05	0.716 ± 0.04	0.747 ± 0.04
GCL[8]	0.266 ± 0.07	0.601 ± 0.05	0.690 ± 0.03	0.736 ± 0.03	0.760 ± 0.03
PCL[9]	$\mathbf{0.296 \pm 0.07}$	0.609 ± 0.05	0.696 ± 0.05	0.747 ± 0.03	0.769 ± 0.03
PPos	0.241 ± 0.10	$\mathbf{0.614 \pm 0.05}$	$\mathbf{0.705 \pm 0.04}$	$\mathbf{0.750 \pm 0.03}$	$\mathbf{0.771 \pm 0.03}$

Note: the evaluation metric is the mean and standard deviation of dice coefficient; M refers to the number of training samples used to finetune the model.

Table 3 Comparison of different methods pretraining on LargeThick and finetuning on SmallThin

Method	10%			50%			100%		
	Dice	P	R	Dice	P	R	Dice	P	R
Random	0.187 6	0.239 3	0.212 3	0.544 6	0.599 7	0.632 8	0.590 6	0.612 7	**0.670 0**
Genesis[22-23]	0.235 5	0.294 5	0.249 0	0.553 9	0.704 5	0.578 1	0.596 5	0.594 4	0.650 1
GCL[8]	0.171 0	0.226 9	0.175 7	0.552 0	0.609 0	0.635 1	0.599 3	0.649 2	0.651 6
PCL[9]	0.200 0	**0.454 2**	0.205 8	0.530 2	0.626 6	0.554 6	0.595 8	0.650 0	0.639 4
PPos	**0.270 0**	0.325 7	**0.302 5**	**0.627 5**	**0.706 7**	**0.679 1**	**0.631 8**	**0.698 1**	0.667 6

Note: 10%, 50%, and 100% refer to the percentage of training samples when the model is finetuned; Dice, P, and R refer to dice coefficient, average voxel precision, and average voxel recall respectively.

Table 4 shows the results for models pretrained on CHD and finetuned on MMWHS. There is a substantial improvement especially when the number of training samples is limited for downstream tasks, which demonstrates the power of direct utilization of positional information in transfer learning.

3.2.3 *Training Overhead*

In Table 5, we compare the GPU memory and training duration that each self-supervised method entails during pretraining on CHD dataset. It is observed that PPos needs the least GPU memory and training duration than other methods. Thus, PPos is more resource-conserving to pretrain while achieving promising results for downstream tasks. Furthermore, the simplicity and lightweight of PPos make it practicable to be viewed as an auxiliary task and combined with other tasks as multi-task learning at supervised learning stage, which we will study and validate in our future work.

Table 4 Comparison of different methods pretraining on CHD and finetuning on MMWHS

Method	$M=2$	$M=4$	$M=6$	$M=10$	$M=16$
Random	0.510 ± 0.16	0.798 ± 0.05	0.842 ± 0.03	0.881 ± 0.02	0.897 ± 0.01
Rotation[35]	0.401 ± 0.20	0.742 ± 0.09	0.808 ± 0.04	0.861 ± 0.02	0.879 ± 0.02
MoCo[5]	0.410 ± 0.19	0.749 ± 0.07	0.811 ± 0.04	0.867 ± 0.02	0.887 ± 0.01
PIRL[36]	0.419 ± 0.18	0.763 ± 0.08	0.823 ± 0.03	0.874 ± 0.02	0.892 ± 0.01
SimCLR[4]	0.426 ± 0.19	0.757 ± 0.07	0.816 ± 0.04	0.870 ± 0.02	0.890 ± 0.01
GCL[8]	0.548 ± 0.16	0.822 ± 0.03	0.848 ± 0.03	0.879 ± 0.02	0.895 ± 0.01
PCL[9]	0.586 ± 0.12	0.829 ± 0.04	0.854 ± 0.03	0.883 ± 0.02	0.898 ± 0.01
PPos	$\mathbf{0.611 \pm 0.13}$	$\mathbf{0.852 \pm 0.05}$	$\mathbf{0.870 \pm 0.02}$	$\mathbf{0.896 \pm 0.01}$	$\mathbf{0.906 \pm 0.01}$

Note: the evaluation metric is the mean and standard deviation of dice coefficient; M refers to the number of training samples used to finetune the model.

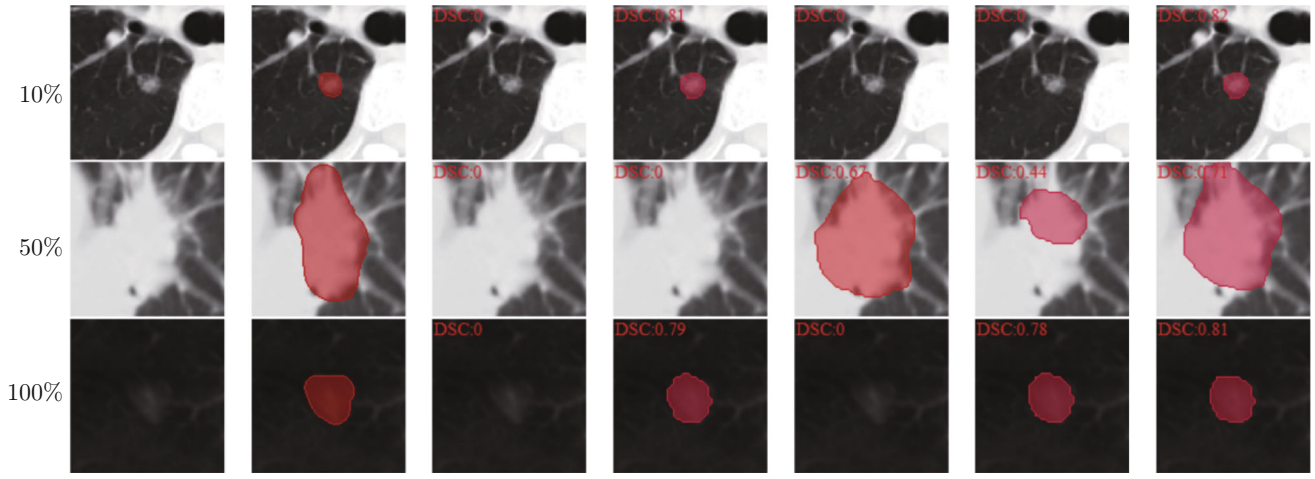
Table 5 Comparison of GPU memory and training duration when pretraining on CHD dataset

Method	GPU memory/GB	Training duration/h
SimCLR[4]	22.7	19.1
GCL[8]	22.7	19.8
PCL[9]	22.7	18.5
PPos	**13.9**	**13.0**

Note: all the methods are pretrained with batch size 32 for 200 epochs on two NVIDIA GeForce GTX 2080Ti GPUs.

3.2.4 *Qualitative Comparison*

Segmentation results of pulmonary nodules in LargeThick dataset are visualized in Fig. 2. It is observed that PPos performs well on hard cases with nodules that are either tiny, attached to adjacent tissue, or visually insignificant, whereas other models may overlook these nodules. It indicates that the relationship between positional information and anatomical patterns is favourable for finding the anomalies such as lung nodules in medical images.



(a) Input (b) Groundtruth (c) Random (d) Genesis (e) GCL (f) PCL (g) PPos

Fig. 2 Visualization of three samples on LargeThick test set (These masks are produced by models finetuned on 10%, 50%, and 100% of training samples on LargeThick dataset)

4 Conclusion

In this paper, we propose a novel and simple self-supervised learning method called PPos. By using PPos to pretrain DCNNs on unlabeled data, promising results can be obtained after finetuning for downstream tasks. Experiments demonstrate the effectiveness and competence of PPos compared with other self-supervised learning methods on four medical image segmentation datasets. It proves that positional information is strong supervision for volumetric medical image segmentation. We hope to extend this method to 3D settings such as 3D spatial coordinates of 3D patches and develop new methods of exploiting positional information in medical images.

References

[1] TAGHANAKI S A, ABHISHEK K, COHEN J P, et al. Deep semantic segmentation of natural and medical images: A review [J]. *Artificial Intelligence Review*,

2021, **54**(1): 137-178.

[2] ZHANG S, XU J C, CHEN Y C, et al. Revisiting 3D context modeling with supervised pre-training for universal lesion detection in CT slices [M]//Medical image computing and computer assisted intervention — MICCAI 2020. Cham: Springer, 2020: 542-551.

[3] JING L L, TIAN Y L. Self-supervised visual feature learning with deep neural networks: A survey [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021, **43**(11): 4037-4058.

[4] CHEN T, KORNBLITH S, NOROUZI M, et al. A simple framework for contrastive learning of visual representations [C]//*37th International Conference on Machine Learning*. Vienna: IMLS, 2020: 1597-1607.

[5] HE K M, FAN H Q, WU Y X, et al. Momentum contrast for unsupervised visual representation learning [C]//*2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Seattle: IEEE, 2020: 9726-9735.

[6] GRILL J B, STRUB F, ALTCHÉ F, et al. Bootstrap your own latent: A new approach to self-supervised learning [C]//*34th Conference on Neural Information Processing Systems*. Vancouver: NIPS, 2020: 21271-21284.

[7] WU Z R, XIONG Y J, YU S X, et al. Unsupervised feature learning via non-parametric instance discrimination [C]//*2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City: IEEE, 2018: 3733-3742.

[8] Chaitanya K, Erdil E, Karani N, et al. Contrastive learning of global and local features for medical image segmentation with limited annotations [C]//*34th Conference on Neural Information Processing Systems*. Vancouver: NIPS, 2020: 12546-12558.

[9] ZENG D W, WU Y W, HU X R, et al. Positional contrastive learning for volumetric medical image segmentation [M]//Medical image computing and computer assisted intervention — MICCAI 2021. Cham: Springer, 2021: 221-230.

[10] RONNEBERGER O, FISCHER P, BROX T. U-Net: Convolutional networks for biomedical image segmentation [M]//Medical image computing and computer-assisted intervention — MICCAI 2015. Cham: Springer, 2015: 234-241.

[11] MILLETARI F, NAVAB N, AHMADI S A. V-Net: Fully convolutional neural networks for volumetric medical image segmentation [C]//*2016 Fourth International Conference on 3D Vision*. Stanford: IEEE, 2016: 565-571.

[12] ÇIÇEK Ö, ABDULKADIR A, LIENKAMP S S, et al. 3D U-net: Learning dense volumetric segmentation from sparse annotation [M]//Medical image computing and computer-assisted intervention — MICCAI 2016. Cham: Springer, 2016: 424-432.

[13] LOU A, GUAN S, LOEW M. DC-UNet: Rethinking the U-Net architecture with dual channel efficient CNN for medical image segmentation [C]//*Medical Imaging 2021: Image Processing*. Online: SPIE, 2021, 11596: 758-768.

[14] ZHOU Z W, SIDDIQUEE M M R, TAJBAKHSH N, et al. UNet: Redesigning skip connections to exploit multiscale features in image segmentation [J]. *IEEE Transactions on Medical Imaging*, 2020, **39**(6): 1856-1867.

[15] ISENSEE F, JAEGER P F, KOHL S A A, et al. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation [J]. *Nature Methods*, 2021, **18**(2): 203-211.

[16] NOROOZI M, FAVARO P. Unsupervised learning of visual representations by solving jigsaw puzzles [M]//Computer vision — ECCV 2016. Cham: Springer, 2016: 69-84.

[17] DOERSCH C, GUPTA A, EFROS A A. Unsupervised visual representation learning by context prediction [C]//*2015 IEEE International Conference on Computer Vision*. Santiago: IEEE, 2015: 1422-1430.

[18] ZHANG R, ISOLA P, EFROS A A. Colorful image colorization [M]//Computer vision — ECCV 2016. Cham: Springer International Publishing, 2016: 649-666.

[19] PATHAK D, KRÄHENBÜHL P, DONAHUE J, et al. Context encoders: Feature learning by inpainting [C]//*2016 IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas: IEEE, 2016: 2536-2544.

[20] KHOSLA P, TETERWAK P, WANG C, et al. Supervised contrastive learning [C]// *34th Conference on Neural Information Processing Systems*. Vancouver: NIPS, 2020: 18661-18673.

[21] CHEN X L, HE K M. Exploring simple Siamese representation learning [C]//*2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Nashville: IEEE, 2021: 15745-15753.

[22] ZHOU Z W, SODHA V, RAHMAN SIDDIQUEE M M, et al. Models genesis: generic autodidactic models for 3D medical image analysis [M]//Medical image computing and computer assisted intervention — MICCAI 2019. Cham: Springer, 2019: 384-393.

[23] ZHOU Z W, SODHA V, PANG J X, et al. Models genesis [J]. *Medical Image Analysis*, 2021, **67**: 101840.

[24] ZHUANG X R, LI Y X, HU Y F, et al. Self-supervised feature learning for 3D medical images by playing a rubik's cube [M]//Medical image computing and computer assisted intervention — MICCAI 2019. Cham: Springer, 2019: 420-428.

[25] ZHU J W, LI Y X, HU Y F, et al. Rubik's Cube+: A self-supervised feature learning framework for 3D medical image analysis [J]. *Medical Image Analysis*, 2020, **64**: 101746.

[26] HAGHIGHI F, TAHER M R H, ZHOU Z W, et al. Transferable visual words: Exploiting the semantics of anatomical patterns for self-supervised learning [J]. *IEEE Transactions on Medical Imaging*, 2021, **40**(10): 2857-2868.

[27] YAN K, LU L, SUMMERS R M. Unsupervised body part regression via spatially self-ordering convolutional neural networks [C]//*2018 IEEE 15th International Symposium on Biomedical Imaging*. Washington: IEEE, 2018: 1022-1025.

[28] LI Z H, ZHANG S, ZHANG J G, et al. MVP-net: Multi-view FPN with position-aware attention for deep universal lesion detection [M]//Medical image computing and computer assisted intervention — MICCAI 2019. Cham: Springer, 2019: 13-21.

[29] XU X W, WANG T C, SHI Y Y, et al. Whole heart and great vessel segmentation in congenital heart disease using deep neural networks and graph matching [M]//Medical image computing and computer assisted intervention — MICCAI 2019. Cham: Springer, 2019: 477-485.

[30] ZHUANG X H. Challenges and methodologies of fully automatic whole heart segmentation: A review [J]. *Journal of Healthcare Engineering*, 2013, **4**(3): 371-408.

[31] ZHUANG X H, SHEN J. Multi-scale patch and multi-modality atlases for whole heart segmentation of MRI [J]. *Medical Image Analysis*, 2016, **31**: 77-87.

[32] QIN X B, ZHANG Z C, HUANG C Y, et al. U^2-Net: Going deeper with nested U-structure for salient object detection [J]. *Pattern Recognition*, 2020, **106**: 107404.

[33] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition [C]//*2016 IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas: IEEE, 2016: 770-778.

[34] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020, **42**(2): 318-327.

[35] GIDARIS S, SINGH P, KOMODAKIS N. Unsupervised representation learning by predicting image rotations [C]//*6th International Conference on Learning Representations*. Vancouver: ICLR, 2018: 1-16.

[36] MISRA I, VAN DER MAATEN L. Self-supervised learning of pretext-invariant representations [C]//*2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Seattle: IEEE, 2020: 6706-6716.