Spiral Contrastive Learning: An Efficient 3D Representation Learning Method for Unannotated CT Lesions

Penghua Zhai[1, 2], **Enwei Zhu**[2], **Baolian Qi**[2], **Xin Wei**[2], and **Jinpeng Li***[1, 2]

[1]HwaMei Hospital, University of Chinese Academy of Sciences (UCAS), Ningbo, China
[2]Ningbo Institute of Life and Health Industry, UCAS, Ningbo, China

***Abstract*—Computed tomography (CT) samples with pathological annotations are difficult to obtain. As a result, the computer-aided diagnosis (CAD) algorithms are trained on small datasets (e.g., LIDC-IDRI with 1,018 samples), limiting their accuracies and reliability. In the past five years, several works have tailored for unsupervised representations of CT lesions via two-dimensional (2D) and three-dimensional (3D) self-supervised learning (SSL) algorithms. The 2D algorithms have difficulty capturing 3D information, and existing 3D algorithms are computationally heavy. Light-weight 3D SSL remains the boundary to explore. In this paper, we propose the spiral contrastive learning (SCL), which yields 3D representations in a computationally efficient manner. SCL first transforms 3D lesions to the 2D plane using an information-preserving spiral transformation, and then learn transformation-invariant features using 2D contrastive learning. For the augmentation, we consider natural image augmentations and medical image augmentations. We evaluate SCL by training a classification head upon the embedding layer. Experimental results show that SCL achieves state-of-the-art accuracy on LIDC-IDRI (89.72%), LNDb (82.09%) and TianChi (90.16%) for unsupervised representation learning. With 10% annotated data for fine-tune, the performance of SCL is comparable to that of supervised learning algorithms (85.75% vs. 85.03% on LIDC-IDRI, 78.20% vs. 73.44% on LNDb and 87.85% vs. 83.34% on TianChi, respectively). Meanwhile, SCL reduces the computational effort by 66.98% compared to other 3D SSL algorithms, demonstrating the efficiency of the proposed method in unsupervised pre-training.**

***Index Terms*—Spiral contrastive learning, 3D representation learning, Computed tomography**

I. Introduction

With the renaissance of deep learning, computer-aided diagnosis (CAD) systems for computed tomography (CT) have achieved many successful applications [1]–[3]. However, Existing CAD systems are mostly developed based on supervised learning which heavily rely on lesion-level annotations. Meanwhile, these annotations are often scarce because only expert experienced radiologists can annotate the lesions. This drives our interest to the unsupervised representation learning. Recent studies show that self-supervised learning (SSL) is an effective approach for learning representations. In the past five years, SSL methods are mostly tailored for unsupervised representations via two-dimensional (2D) images [4]–[7]. However, it is difficulty for 2D SSL algorithms to capture three-dimensional (3D) information of CT scans. Therefore, several 3D SSL are explored to learn the 3D representations of lesions [8], [9]. *Nevertheless, existing 3D SSL algorithms suffer from the computationally heavy and slow convergence problems*.

To address these issues, we propose a novel light-weight spiral contrastive learning (SCL) method to efficiently learn 3D representations for unannotated CT lesions with fewer computational resources and parameters. Unlike existing 3D SSL methods, we feed a transformed 2D view into the network, which is converted from a 3D lesion volume by the spiral transformation [10]. To fully learn the 3D representations and spatial information, the spiral transformation has the following property: The 2D views should preserve the 3D features of lesions as much as possible. Therefore, the proposed SCL is able to learn the 3D representations by the spiral transformation. We regard the views of the same lesion as positive pairs, and the views of different lesions as negative pairs. We learn representations by minimizing the contrastive loss to obtain a embedding space with the property of within-lesion compactness and between-lesion separability. We validate the effectiveness of our method on three lung CT datasets. The experimental results show that the proposed SCL outperforms state-of-the-art SSL methods.

The contributions of this paper are as follows: 1) An information-preserving spiral transformation is introduced to convert each 3D lesion to a 2D view. The 2D view can preserve the correlation between 3D lesion adjacent pixels and retain the spatial relationship of texture features for the 3D volume. 2) We propose a novel spiral contrastive learning (SCL) method, which yields 3D transformation-invariant representations in a computionally efficient manner. It is a light-weight 3D SSL algorithm. 3) We evaluate the proposed SCL by training a simple classification head upon the embedding layer. Experimental results show that SCL achieves state-of-the-art accuracy on LIDC-IDRI, LNDb and TianChi for unsupervised representation learning. With 10% annotated data for fine-tuning, the performance of SCL is comparable to that of supervised learning algorithms. Meanwhile, SCL reduces the computational effort by 66.98% compared to other 3D SSL algoithms, demonstrating the efficiency of the proposed method in unsupervised pre-training.

This work was supported by National Natural Science Foundation of China (62106248), and Ningbo Natural Science Foundation (202003N4270).

*Corresponding author: Jinpeng Li (lijinpeng@ucas.ac.cn).

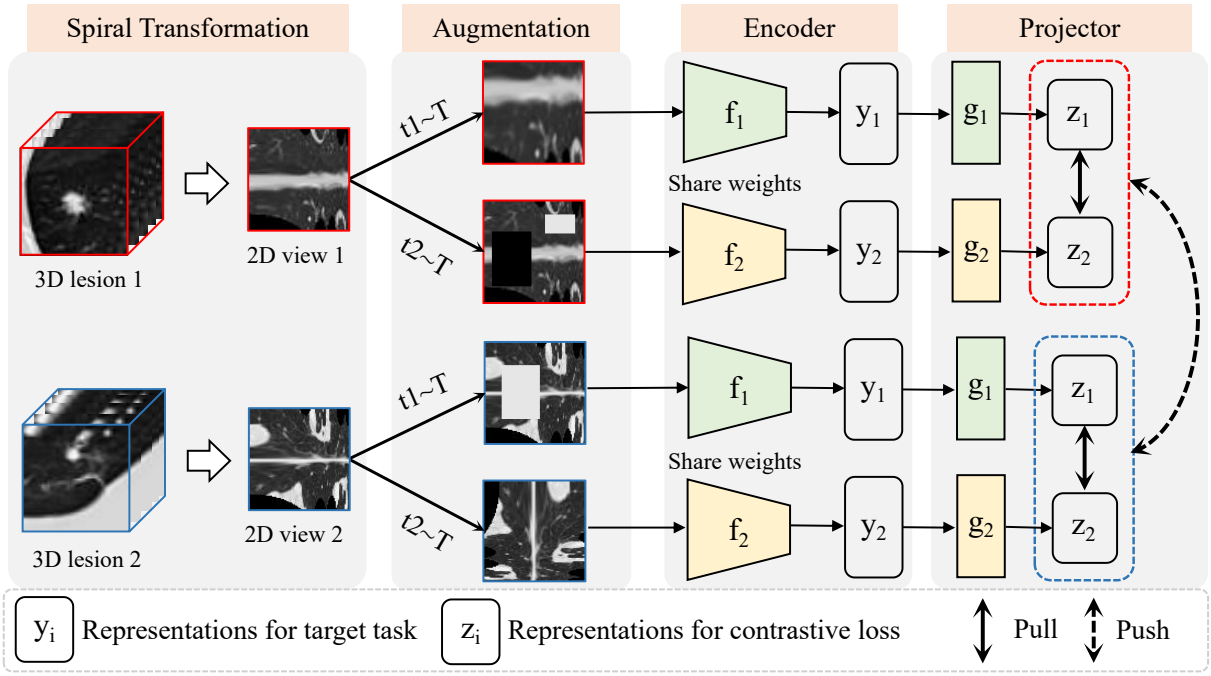

Fig. 1. Illustration of the training stage of the SCL. Each lesion volume is converted to a 2D view by the information-preserving spiral transformation, and two augmented views are generated by the augmentation $\boldsymbol{t}_i$, where $i = 1, 2$. Then, we construct a light-weight contrastive learning to generate transformation-invariant representations for each augmented view. We make views of the same lesions attract each other and views of different lesions repeal each other to learn discriminative features. Being optimized by a contrastive loss, the shared embedding space is endowed with well local-aggregating properties and the spread-out properties of representations are preserved. $\boldsymbol{y}_i$ and $\boldsymbol{z}_i$represent the features learned from encoder f_i and g_i

II. Related Works

A. Self-supervised Learning

SSL, especially contrastive learning, is an extensively-studied area in recent years due to its effectiveness in unsupervised representation learning [11], [12]. Gidaris et al. [13] learned image features by recognizing the rotation of the input image, and they proved that this simple task actually provides a very powerful supervision signal for semantic feature learning. He et al. [6] facilitated contrastive learning by building a dynamic dictionary with a queue and a moving-averaged encoder. Chen et al. [5] proposed a simple contrastive visual representation learning framework by aggregating the views of the same image and separating the other views. Furthermore, a series of transformation-based methods [14], [15] and pretext task [4], [16] are proposed to achieve SSL.

SSL methods are also used to achieve medical image representation learning [8], [17], [18]. Chen et al. [19] proposed a novel context restoration method for fetal 2D ultrasound detection to learn semantic image features. However, the 2D network has difficulty in capturing 3D information of lesions. Zhou et al. [8] learned the common anatomical representation automatically by using the sophisticated yet recurrent anatomy in medical images as supervision signals. Zhu et al. [9] proposed a pretext task, i.e., Rubik's cube+, to force networks to learn translation and rotation invariant features from the 3D medical data. Li et al. [17] presented a novel multi-modal self-supervised feature learning method by capturing the semantically shared information across different modalities and the apparent visual similarity between patients for retinal disease diagnosis. However, existing 3D SSL methods are computationally heavy, such as the number of parameters of Rubik's cube+ [9] is $84M$, while the parameters of SimCLR [5] and MoCo [6] are less than $12M$. 2D contrastive learning has fewer 66.98% parameters than 3D contrastive learning with same backbone (such as ResNet18 and 3D ResNet18 [20]), the 2D. It is necessary to further explore the SSL by combining the advantages of 2D and 3D networks to learn spatial information of CT scans.

B. Automatic Lung CT Diagnostic Models

CAD systems are used to assist radiologists in interpreting disease-related information in medical images [1], [2]. Considering the 3D characteristics of lesions, Setio et al. [21] achieved the lung nodule diagnosis by composing the 3D lesion volume into nine views. Although the method achieved good results by fusing multiple 2D views, the multi-view method is not always suitable for radiologists. In the process of CT screening, radiologists usually observe the complete 3D lesions in CT scans for diagnosis. Therefore, Mei et al. [22] proposed a 3D slice-aware network (SANet) for pulmonary nodule detection and introduced a multi-scale feature maps to reduce the false positive. Shen et al. [23] presented a novel interpretable deep hierarchical semantic CNN by combining low-level and high-level features to predict malignant nodules. However, compared to models trained on natural images,

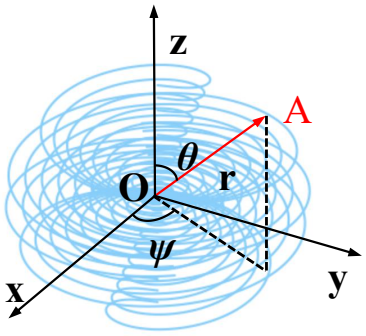

Fig. 2. Coordinate system of spiral transformation. The coordinate origin O is the center-point of spiral transformation. And the spiral line is calculated by Θ, Ψ and r.

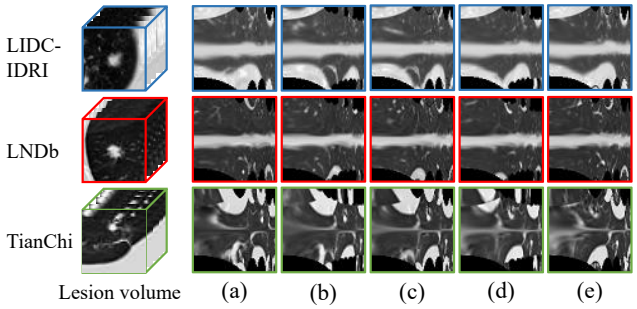

Fig. 3. The views generated by the spiral transformation in different coordinate systems for each dataset. For each column, the transformed 2D views are generated from the same 3D lesion. The rows from (b) to (e) show that the coordinate system is rotated by 50, 90, 140 and 180 degrees relative to (a) in the x-y plane, respectively.

the medical models are trained on the small datasets (e.g. LIDC-IDRI [24]) with limited accuracy and generalization capabilities [25]. It is a challenge to achieve the high-quality 3D CAD systems with limited labeled samples.

III. Methods

In this section, we illustrate the novel light-weight spiral contrastive learning (SCL) method as shown in Fig. 1. Assuming that the suspicious lesions have been detected, and we first convert each 3D lesion volume to a 2D view by the spiral transformation. We then learn 3D transformation-invariant representations with a computationally efficient manner. We finally evaluate the representations by training a simple classification head.

A. Information-preserving Spiral Transformation

Typically, 3D networks are better than 2D networks for learning spatial information of lesions, while 3D networks are computationally heavy. Inspired by the SpiralTransform [10], we introduce a information-preserving spiral transformation to convert each 3D lesion to a 2D view, as shown in Fig. 2. Therefore, a light-weight contrastive learning method can be developed to obtain the 3D representations. During the transformation, the method preserves the correlation between original adjacent pixels. Therefore, the 2D view can retain the spatial relationship of texture features from the 3D lesion volume.

We denote the dataset by $S = \{(X_i, y_i)\}_{i=1}^{N}$, where $X_i \in \mathbb{R}^{d\times w\times h}$ is the i-th 3D lesion volume, y_i denotes the label of X_i, and the N is the number of samples. We select the center point of the lesion volume $X_i \in \mathbb{R}^{d\times w\times h}$ as the midpoint O of the spiral transformation, and the space rectangular coordinate system is established with midpoint O. The maximum radius R of the transformation depends on the maximum distance from the tumor edge to the O point. The spiral line A is determined by an azimuth angle Ψ, an elevation angle $1-\Theta$, and the distance to O in the 3D space. The coordinates of A can be expressed as:

$$\begin{cases} x = r\sin\Theta\cos\Psi \\ y = r\sin\Theta\sin\Psi \\ z = r\cos\Theta \end{cases} \text{where} \begin{cases} 0 \le \Theta \le \pi/2 \\ 0 \le \Psi \le 2\pi \\ -R \le r \le R \end{cases}, \quad (1)$$

Let the circle on the equator have $2M$ sampling points, and the arc length d of the adjacent sampling points is defined as the distance between two points on the equator [26]. Therefore, the number of sampling points in a horizontal plane corresponding to angle Θ is:

$$2\pi r|sin\Theta|/d = 2\pi r|sin\Theta|/(\pi r/M) = 2M|sin\Theta|, \quad (2)$$

where the Θ is divided into M angles evenly within the value range. In the case of a specified radius, if M is large enough, the total number of sampling points on a spiral line can be calculated as:

$$\begin{aligned} m &= \int_0^M 2M\sin\frac{k\pi}{2M}dk = 2M\int_0^M \frac{2M}{\pi}\sin\frac{k\pi}{2M}d\frac{k\pi}{2M} \\ &= \frac{4M^2}{\pi}\int_0^{\pi/2}\sin x dx = \frac{4M^2}{\pi}, \end{aligned} \quad (3)$$

The size of the transformed 2D view is $2R \times m$ due to the number of sampling points is same for each r from (3). We set the R to 46, and the M to 12 in this paper. The gray value of the sampling points is calculated by trilinear interpolation, whose coordinates in the 3D space are then mapped to the original matrix position.

The result of spiral transformation depends on the spatial rectangular coordinate system and the parameters of the transformation [10]. For the same 3D data, once the direction of the coordinate system in the spiral transformation is modified, a new transformed image can be obtained. Fig. 3 simulates the data generation results using the spiral transformation in different directions.

B. Light-weight SSL Representation Learning

Contrastive learning aims to construct a latent embedding space that separates samples from different clusters in an unsupervised manner. The 2D algorithms have difficulty capturing 3D information, and existing 3D algorithms are computationally heavy. Therefore, we design a light-weight contrastive learning to achieve 3D lesion representation learning. In this section, we describe the process of improving the similarity within lesions and separability between lesions by a contrastive loss. As shown in Fig. 1. The two networks have the same

structure and share weights. Encoder $f(\cdot)$ is used to learn embedding space, and projector $g(\cdot)$ is used to eliminate the semantically irrelevant low-level information from the representation.

Given an input volume X, a 2D view $\boldsymbol{x}$ is generated by the spiral transformation. Then, two views $\boldsymbol{v}_1 = \boldsymbol{t}1(\boldsymbol{x})$ and $\boldsymbol{v}_2 = \boldsymbol{t}_2(\boldsymbol{x})$ are generated by two different sets of augmentations $\boldsymbol{t}1$ and $\boldsymbol{t}2$, and the two set of augmentations are randomly sampled from the distribution $\boldsymbol{T}$. The two transformed 2D views are treated as the input images for the network. Specifically, the distribution $\boldsymbol{T}$ is divided into two categories: medical image augmentation (MIA) and natural image augmentation (NIA). The NIA combines with standard data augmentation methods such as cropping, resizing, flipping, and Gaussian blur [5]. The MIA includes in-painting, out-painting, local pixel shuffling and non-linear transformation [8].

The aim of the encoder $f(\cdot)$ is to extract representation vectors from augmented samples. The representation $\boldsymbol{y}_i = f(\boldsymbol{x}_i)$, where $\boldsymbol{y}_i \in \mathbb{R}^d$ obtained from the output of the average pooling layer. Then, a projection head $g(\cdot)$ maps the representation to the space where contrastive loss is applied. We use a multi-layer perceptron (MLP) with two hidden layers as the projection head to obtain $\boldsymbol{z}_i = g(\boldsymbol{y}_i) = g(f(\boldsymbol{x}_i))$. We define the contrastive loss on $\boldsymbol{z}_i$ rather than $\boldsymbol{y}_i$. The importance of using the representation before the nonlinear projection is due to loss of information induced by the contrastive loss. The $\boldsymbol{z}_i = g(\boldsymbol{y}_i)$ is trained to be invariant to data transformation. Thus, $g(\cdot)$ can remove information that may be useful for the downstream task, such as the color information. More information can be maintained in $\boldsymbol{y}_i$ by using the nonlinear transformation $g(\cdot)$ [5]. The projector $g(\cdot)$ includes a linear layer with an output size of 512, a batch normalization layer, a ReLU activation function, and a linear layer with an output size of 128.

Finally, we minimize the contrastive loss by aggregating the positive pair and separating the negative pair. Considering N samples in a batch, there are $2N$ views augmented by the $\boldsymbol{t}1$ and $\boldsymbol{t}2$. We treat the two augmented views from the same lesion as positive pair, the other $2(N-1)$ augmented views as negative pair. We use the cosine similarity to represent the similarity of the representations of different views:

$$\mathrm{sim}(\boldsymbol{p}_1, \boldsymbol{p}_1) = \boldsymbol{p}_1^\top \boldsymbol{p}_2 / \|\boldsymbol{p}_1\| \|\boldsymbol{p}_2\| \tag{4}$$

where $\boldsymbol{p}_1, \boldsymbol{p}_2$ denote the normalized features of $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ from the projection head, respectively. Then, the loss function for a positive pair of examples (i, j) is defined as:

$$\ell_{i,j} = -\log \frac{\exp\left(\mathrm{sim}\left(\boldsymbol{z}_i, \boldsymbol{z}_j\right)/\tau\right)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp\left(\mathrm{sim}\left(\boldsymbol{z}_i, \boldsymbol{z}_k\right)/\tau\right)} \tag{5}$$

where $\mathbb{1}_{[k \neq i]} \in \{0, 1\}$ is an indicator function evaluating to 1 if $k \neq i$ and τ denotes a temperature parameter. The final loss is computed across all positive pairs in a batch, both (i, j) and (j, i).

C. 3D Lesion Diagnosis Task

After the pre-training of the SCL, we evaluate the encoder by a lesion diagnosis task, assuming that the detection of suspicious lesions has been completed. A classification head $p(\cdot)$ is designed to achieve the task. The classification head only includes a batch normalizatin layer and a fully connected layer, and the input of $p(\cdot)$ is the output of the encoder $f(\cdot)$ rather than the projector $g(\cdot)$. We apply the SCL to lung CT dataset 10 times independently, with the 10-fold cross validation. The performance is assessed by the mean and standard deviation of accuracy and area under the receiver operator curve (AUC) [25], [30]. Accuracy shows the the proposed model can correctly distinguish between malignant and benign nodules and defines as :

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{6}$$

where TP is true positive; TN is true negative; FP is false positive; FN is false negative. AUC takes into account the sensitivity and specificity in a comprehensive manner.

IV. Experiments and Results

In order to evaluate the proposed SCL, we conduct a series of experiments on lung CT datasets and compared SCL with state-of-the-art SSL models.

A. Datasets

1) LIDC-IDRI: The LIDC-IDRI [1] is a commonly-used CT dataset, which contains 1,018 CT scans for lung nodule diagnosis [24]. The malignancy of each nodule is evaluated with a 5-point scale from benign to malignant by up to four experienced radiologists in two stages. In this study, scans with thickness thicker than $2.5mm$ are eliminated as this could easily lead to the omission of small nodules [21]. We calculate the mean malignancy (l) of a nodule which is annotated by at least three radiologists, and annotate a nodule whose $l < 3$ as benign, a nodule whose $l = 3$ as uncertain and a nodule whose $l > 3$ as malignant. To reduce the impact of uncertain evaluation, we exclude all uncertain lung nodules from the dataset. Finally, there are totally 369 benign and 335 malignant nodules.

2) LNDb: The LNDb [2] dataset contains a total of 294 CT scans [31]. Each CT scan is read by at least one radiologist with a 5-point scale from benign to malignant. We process the LNDb in the LIDC-IDRI way, with the difference that we calculate the mean malignancy (l) of all annotated nodules, since few nodules annotated by at least three radiologists. Finally, there are 451 benign and 768 malignant nodules.

3) TianChi: The TianChi [3] lung disease diagnosis dataset contains a total of 1,470 CT scans with four diseases. The annotation process is divided into two stages: two physicians perform the original annotation, then a third independent

[1] https://wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI

[2] https://lndb.grand-challenge.org/Data/

[3] https://tianchi.aliyun.com/competition/entrance/231724/information?lang=en-us

TABLE I
COMPARISON OF SCL WITH STATE-OF-THE-ART SSL METHODS ON THE THREE DATASETS ($Mean(\%) \pm std(\%)$)

	LIDC-IDRI		LNDb		TianChi
	AUC	**Accuracy**	**AUC**	**Accuracy**	**Accuracy**
Natural Image Augmentation					
Context [27]	64.93 ± 1.60	56.88 ± 0.97	64.57 ± 0.32	62.50 ± 0.58	64.13 ± 1.05
RotNet [13]	64.96 ± 1.14	56.61 ± 2.21	63.80 ± 0.59	58.40 ± 0.68	60.22 ± 0.39
MoCo [6]	71.07 ± 0.11	71.29 ± 0.13	66.09 ± 0.48	67.97 ± 0.46	73.44 ± 0.67
MoCo V2 [15]	71.88 ± 0.33	71.83 ± 0.25	68.20 ± 0.26	70.31 ± 0.44	76.13 ± 0.72
SimCLR [5]	78.58 ± 0.52	79.04 ± 0.50	69.71 ± 0.69	73.23 ± 0.91	79.08 ± 0.56
BYOL [28]	78.35 ± 0.22	65.19 ± 0.51	63.12 ± 0.33	67.97 ± 0.70	63.04 ± 0.85
SimSiam [29]	76.39 ± 0.93	69.11 ± 1.12	68.20 ± 0.44	71.88 ± 0.46	73.93 ± 0.43
Medical Image Augmentation					
MoCo [6]	73.27 ± 0.55	74.15 ± 0.53	70.44 ± 0.15	70.02 ± 0.21	75.63 ± 0.34
MoCo V2 [15]	78.45 ± 0.85	78.69 ± 0.55	70.30 ± 0.23	70.35 ± 0.35	77.13 ± 0.65
SimCLR [5]	79.98 ± 0.50	80.31 ± 0.71	71.21 ± 0.96	75.00 ± 1.47	80.72 ± 0.93
BYOL [28]	70.49 ± 0.18	65.77 ± 0.94	69.06 ± 0.26	71.87 ± 0.31	69.27 ± 0.33
SimSiam [29]	77.10 ± 1.12	70.32 ± 1.69	72.30 ± 0.36	74.72 ± 0.27	75.38 ± 0.59
Models Genesis [8]	76.29 ± 2.33	65.83 ± 1.56	61.73 ± 0.34	61.94 ± 0.30	63.93 ± 0.43
Rubik's Cube+ [9]	82.07 ± 0.44	81.21 ± 0.16	75.63 ± 0.51	66.67 ± 0.39	76.14 ± 0.73
Restoration [19]	85.60 ± 0.31	78.75 ± 0.84	74.09 ± 0.51	67.71 ± 0.60	79.27 ± 0.95
Ours					
SCL (Ours)	$\mathbf{93.89} \pm 0.60$	$\mathbf{89.72} \pm 0.38$	$\mathbf{77.95} \pm 0.78$	$\mathbf{82.09} \pm 0.67$	$\mathbf{90.16} \pm 0.02$

physician perform the disambiguation to ensure the consistency of the data annotation. There are 3,264 nodules, 3,613 Streak shadows, 4,201 arteriosclerosis or calcification and 1,140 lymph node calcification.

B. Implementation Details

We transform each 3D lesion to a corresponding 2D view, and resize each view to 224×224. To generate the different transformed 2D views, we rotate the coordinate systems clockwise with angles of $50°$, $90°$, $140°$ and $180°$. We use the ResNet18 [20] as the backbone without fully connected layer. Adam [32] with a base learning rate of 0.001 is used to optimize the SCL. The training epoch is set to 1000 with an early-stopping of 50. We set the batch size to 64, and the weight decay rate is $1e-4$. The temperature τ is set to 0.07. All the experiments are conducted with PyTorch [4] using a single Nvidia RTX 2080 ti 11GB GPU.

C. Lung Nodule Classification

We train a simple classifier on features extracted from the frozen pre-trained SCL to evaluate the representations, as shown in Table I. The Models Genesis [8], Rubik's Cube+ [9] and Restoration [19] are developed for medical images, and the others are developed based on natural images.

Comparing the existing state-of-the-art SSL methods, the accuracy of the methods based on MIA is higher than the method based on NIA when applying them to CT scans. This proves that MIA is more effective for medical images. For the proposed SCL on LIDC-IDRI, we obtain an accuracy of 89.72%. Compared to SimCLR (MIA) and Rubik's Cube+, SCL achieves 9% and 8% improvement, respectively. Compared to state-of-the-art SSL methods, the highest accuracies are also obtained on both LNDb (82.09%) and TianChi (90.16%). The results show that the proposed SCL has superior performance than previous SSL methods.

[4]https://pytorch.org/

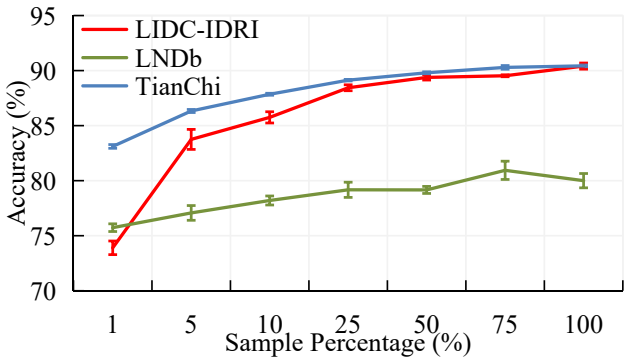

Fig. 4. Results of fine-tuning with different numbers of samples.

D. Fine-tuning on Three Datasets

To further evaluate the effectiveness of the SCL, we fine-tune the model with limited annotated samples in each dataset. As shown in Table II, we fine-tune the SCL with 10% datasets, and the accuracies are 85.75% and 78.20% on LIDC-IDRI and

TABLE II
FINE-TUNING WITH 10% SAMPLES ON THE THREE DATASETS ($Mean(\%) \pm std(\%)$)

	LIDC-IDRI		LNDb		TianChi
	AUC	**Accuracy**	**AUC**	**Accuracy**	**Accuracy**
Supervised	$\mathbf{90.90} \pm 0.47$	85.03 ± 0.50	70.65 ± 0.81	73.44 ± 0.21	83.34 ± 0.33
Supervised (3D)	78.88 ± 0.31	76.60 ± 0.42	$\mathbf{77.43} \pm 0.85$	$\mathbf{80.25} \pm 0.53$	78.13 ± 0.75
SimCLR (NIA) [5]	85.91 ± 0.74	84.90 ± 0.73	73.62 ± 0.35	72.97 ± 0.99	79.08 ± 0.56
SimCLR (MIA) [5]	85.47 ± 0.17	85.46 ± 0.22	73.85 ± 0.92	78.34 ± 0.91	80.72 ± 0.93
SCL (Ours)	90.11 ± 1.15	$\mathbf{85.75} \pm 1.01$	69.16 ± 1.34	78.20 ± 0.63	$\mathbf{87.85} \pm 0.09$

LNDb, respectively, and they are comparable to Supervised. We also conduct multi-disease diagnosis on TianChi, and the diagnostic accuracy of SCL is higher (4%) than that of Supervised. Overall, the SCL fine-tuned with 10% datasets is better than SimCLR and is comparable to the Supervised model.

We also investigate the performance of SCL for fine-tuning with different percentages of samples on the three datasets and summarize the results in Fig. 4. According to the overall trend in Fig. 4, the accuracy increases with the increase of samples. Even though the samples are very small (e.g., percentage=1%), the accuracy of SCL also reaches a relatively higher value. It proves that SCL is able to learn high-quality lesion representations, therefore, only a small number of samples are needed to learn the category of lesions. Therefore, we can construct high-performance CAD systems based on SSL with a small number of samples.

V. CONCLUSION

In this paper, we propose the SCL, a novel spiral contrastive learning framework for yielding 3D representations in a computationally efficient manner. Unlike existing 3D SSL methods, SCL is a light-weight network due to we convert each 3D lesion into a 2D view by the information-preserving spiral transformation. Most studies apply the 2D slices in a transverse plane as the input, so it ignores the correlation of information between the adjacent layers. Spiral transformation takes the whole 3D lesion to a 2D plane, which retains the correlation of the features in 3D space. Experimental results on three CT datasets show that the proposed SCL outperforms state-of-the-art SSL methods, indicating the superiority of SCL in *unsupervised representation learning*. Being fine-tuned with a small percentage of the datasets (10%), our model is comparable to the 3D fully supervised model, demonstrating its superiority in *small-data scenarios* and the potential of reducing the annotation efforts in 3D CAD systems.

The main insufficiency of our work is that SCL is designed for the lesion diagnosis task and is not optimized for the localization task on the whole CT volumes. The effectiveness of SCL methods in lesion localization (often with background dominance) needs to be explored. In the future, we will adapt SCL to more tasks (lesion localization and segmentation) and evaluate it on more imaging modalities such as the magnetic resonance imaging.

REFERENCES

[1] J. Yanase and E. Triantaphyllou, "A systematic survey of computer-aided diagnosis in medicine: Past and present developments," *Expert Systems with Applications*, vol. 138, p. 112 821, 2019.

[2] S. Zhang, F. Sun, N. Wang, C. Zhang, Q. Yu, M. Zhang, P. Babyn, and H. Zhong, "Computer-aided diagnosis (cad) of pulmonary nodule of thoracic ct image using transfer learning," *Journal of digital imaging*, vol. 32, no. 6, pp. 995–1007, 2019.

[3] H.-P. Chan, L. M. Hadjiiski, and R. K. Samala, "Computer-aided diagnosis in the era of deep learning," *Medical physics*, vol. 47, no. 5, e218–e227, 2020.

[4] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *European conference on computer vision*, Springer, 2016, pp. 69–84.

[5] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*, PMLR, 2020, pp. 1597–1607.

[6] 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.

[7] P. Zhai, H. Cong, G. Zhao, C. Fang, J. Li, T. Cai, and H. He, "Mvcnet: Multiview contrastive network for unsupervised representation learning for 3d ct lesions," *arXiv preprint arXiv:2108.07662*, 2021.

[8] Z. Zhou, V. Sodha, M. M. R. Siddiquee, R. Feng, N. Tajbakhsh, M. B. Gotway, and J. Liang, "Models genesis: Generic autodidactic models for 3d medical image analysis," in *International conference on medical image computing and computer-assisted intervention*, Springer, 2019, pp. 384–393.

[9] J. Zhu, Y. Li, Y. Hu, K. Ma, S. K. Zhou, and Y. Zheng, "Rubik's cube+: A self-supervised feature learning framework for 3d medical image analysis," *Medical image analysis*, vol. 64, p. 101 746, 2020.

[10] X. Chen, X. Lin, Q. Shen, and X. Qian, "Combined spiral transformation and model-driven multi-modal deep learning scheme for automatic prediction of tp53

mutation in pancreatic cancer," *IEEE Transactions on Medical Imaging*, vol. 40, no. 2, pp. 735–747, 2020.

[11] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[12] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, and J. Tang, "Self-supervised learning: Generative or contrastive," *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[13] N. Komodakis and S. Gidaris, "Unsupervised representation learning by predicting image rotations," in *International Conference on Learning Representations (ICLR)*, 2018.

[14] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, Springer, 2020, pp. 776–794.

[15] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," *arXiv preprint arXiv:2003.04297*, 2020.

[16] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *European conference on computer vision*, Springer, 2016, pp. 649–666.

[17] X. Li, M. Jia, M. T. Islam, L. Yu, and L. Xing, "Self-supervised feature learning via exploiting multi-modal data for retinal disease diagnosis," *IEEE Transactions on Medical Imaging*, vol. 39, no. 12, pp. 4023–4033, 2020.

[18] J. Li, G. Zhao, Y. Tao, P. Zhai, H. Chen, H. He, and T. Cai, "Multi-task contrastive learning for automatic ct and x-ray diagnosis of covid-19," *Pattern recognition*, vol. 114, p. 107 848, 2021.

[19] L. Chen, P. Bentley, K. Mori, K. Misawa, M. Fujiwara, and D. Rueckert, "Self-supervised learning for medical image analysis using image context restoration," *Medical image analysis*, vol. 58, p. 101 539, 2019.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[21] A. A. A. Setio, F. Ciompi, G. Litjens, P. Gerke, C. Jacobs, S. J. Van Riel, M. M. W. Wille, M. Naqibullah, C. I. Sánchez, and B. Van Ginneken, "Pulmonary nodule detection in ct images: False positive reduction using multi-view convolutional networks," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1160–1169, 2016.

[22] J. Mei, M.-M. Cheng, G. Xu, L.-R. Wan, and H. Zhang, "Sanet: A slice-aware network for pulmonary nodule detection," *IEEE transactions on pattern analysis and machine intelligence*, 2021.

[23] S. Shen, S. X. Han, D. R. Aberle, A. A. Bui, and W. Hsu, "An interpretable deep hierarchical semantic convolutional neural network for lung nodule malignancy classification," *Expert systems with applications*, vol. 128, pp. 84–95, 2019.

[24] S. G. Armato III, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. I. Henschke, E. A. Hoffman, *et al.*, "The lung image database consortium (lidc) and image database resource initiative (idri): A completed reference database of lung nodules on ct scans," *Medical physics*, vol. 38, no. 2, pp. 915–931, 2011.

[25] Y. Xie, Y. Xia, J. Zhang, Y. Song, D. Feng, M. Fulham, and W. Cai, "Knowledge-based collaborative deep learning for benign-malignant lung nodule classification on chest ct," *IEEE transactions on medical imaging*, vol. 38, no. 4, pp. 991–1004, 2018.

[26] J. Wang, R. Engelmann, and Q. Li, "Segmentation of pulmonary nodules in three-dimensional ct images by use of a spiral-scanning technique," *Medical Physics*, vol. 34, no. 12, pp. 4678–4689, 2007.

[27] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1422–1430.

[28] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, *et al.*, "Bootstrap your own latent: A new approach to self-supervised learning," *arXiv preprint arXiv:2006.07733*, 2020.

[29] X. Chen and K. He, "Exploring simple siamese representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 750–15 758.

[30] P. Zhai, Y. Tao, H. Chen, T. Cai, and J. Li, "Multi-task learning for lung nodule classification on chest ct," *IEEE access*, vol. 8, pp. 180 317–180 327, 2020.

[31] J. Pedrosa, G. Aresta, C. Ferreira, M. Rodrigues, P. Leitão, A. S. Carvalho, J. Rebelo, E. Negrão, I. Ramos, A. Cunha, *et al.*, "Lndb: A lung nodule database on computed tomography," *arXiv preprint arXiv:1911.08434*, 2019.

[32] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR (Poster)*, 2015.