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

Penghua Zhai1,2, Enwei Zhu2, Xin Wei2, 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 augmentation and medical image augmentation. 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-tuning, 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—Light-weight contrastive representation learning, 3D medical image analysis, Spiral transformation

1. INTRODUCTION

With the renaissance of deep learning, computer-aided diagnosis (CAD) systems for computed tomography (CT) have achieved many successful applications. However, they heavily rely on lesion-level annotations, which are often scarce because expert experienced radiologists are needed to annotate lesions. This drives our interest to the unsupervised representation learning, of which self-supervised learning (SSL) is one of the most effective approaches. However, most SSL methods are tailored for unsupervised representations via two-dimensional (2D) images [1, 2, 3]. It is difficult 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 [4, 5], which suffer from computationally heavy and slow convergence.

To address these issues, we propose a novel light-weight spiral contrastive learning (SCL) method to efficiently learn 3D representations 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 [6]. The spiral transformation has the following property: The 2D views should preserve the 3D features of lesions as much as

Fig. 1. Illustration of the training stage of the SCL. Each lesion volume is converted to a 2D view by the spiral transformation, and two augmented views are generated by the augmentation. Then, we construct a light-weight contrastive learning framework by making views of the same lesions attract each other and views of different lesions repel each other to learn discriminative representation.
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 an embedding space with the property of within-lesion compactness and between-lesion separability. We validate the effectiveness of SCL on three lung CT datasets. The experimental results show that the SCL outperforms state-of-the-art SSL methods.

The contributions of this paper are as follows: 1) We introduce the spiral transformation to contrastive learning framework. The spiral transformation 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 light-weight spiral contrastive learning method, which yields 3D transformation-invariant representations in a computationally efficient manner. 3) The experimental results show that SCL achieves state-of-the-art accuracy on three datasets for unsupervised representation learning. Meanwhile, SCL reduces the computational effort by 66.98% compared to 3D SSL algorithms, demonstrating the efficiency of the proposed method in unsupervised pre-training.

2. METHODS

In this section, we introduce a novel SCL method to solve the lesion diagnosis task, as shown in Fig. 1. Assuming that the suspicious lesions have been detected, we convert each 3D lesion volume to a 2D view by the spiral transformation.

2.1. Spiral Transformation

As shown in Fig. 2, inspired by the SpiralTransform [6], we introduce a spiral transformation to convert each 3D lesion to a 2D view. We select the center point of the lesion volume as the midpoint $O$ of the spiral transformation, and we denote the maximum radius by $R$. The spiral line $A$ is determined by an azimuth angle $\Psi$, an elevation angle $1 - \Theta$, and the distance to $O$ in the 3D space.

![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 $\Theta$, $\Psi$ and $r$.](image)

The size of the transformed 2D view is $2R \times m$ due to the number of sampling points is same for each $r$ from (1). We set the $R$ to 46, and the $M$ to 12 in this paper.

The result of spiral transformation depends on the spatial rectangular coordinate system and the parameters of the transformation [6]. Fig. 3 simulates the views transformed in different directions.

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

Given an input volume $X$, a 2D view $x$ is generated by the spiral transformation. Then, two views $v_1 = t_2(x)$ and $v_2 = t_2(x)$ are generated by $t1$ and $t2$, which are randomly sampled from the distribution $T$. Specifically, the distribution $T$ is divided into two categories: medical image augmentation (MIA) and natural image augmentation (NIA). The MIA combines cropping, resizing, flipping, and Gaussian blur [1]. The MIA includes in-painting, out-painting, local pixel shuffling.

![Fig. 3. The views generated by the spiral transformation in different coordinate systems. For each row, the 2D views are generated from the same 3D lesion. The columns 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.](image)
and non-linear transformation [4].

The aim of the encoder \( f(\cdot) \) is to extract representation vectors from augmented views. The representation \( y_i = f(x_i) \), where \( y_i \in \mathbb{R}^d \) 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, which is a multi-layer perceptron (MLP) with two hidden layers, \( z_i = g(y_i) = g(f(x_i)) \). The specific network structure of 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. The \( z_i = g(y_i) \) 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 \( y_i \) by using the nonlinear transformation \( g(\cdot) \) [1].

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 \( T \). 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:

\[
\operatorname{sim}(p_1, p_1) = \frac{p_1^T p_2}{\|p_1\| \|p_2\|}
\]

where \( p_1, p_2 \) denote the normalized features of \( v_1 \) and \( v_2 \) from the projection head, respectively. Then, the loss function for a positive pair of examples \((i, j)\) is defined as:

\[
\ell_{i,j} = -\log \frac{\exp(\operatorname{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k\neq i]} \exp(\operatorname{sim}(z_i, z_k)/\tau)}
\]

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

After the pre-training of the SCL, we evaluate the encoder by the lesion diagnosis task, which is achieved by a simple classification head \( p(\cdot) \). The classification head only includes a batch normalization 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).

### 3. EXPERIMENTS AND RESULTS

#### 3.1. Datasets

**LIDC-IDRI:** The LIDC-IDRI is a commonly-used CT dataset, which contains 1,018 CT scans for lung nodule diagnosis [13, 14]. The malignancy of each nodule is evaluated with a 5-point scale from benign to malignant. There are totally 369 benign and 335 malignant nodules.

**LNDb:** The LNDb contains a total of 294 CT scans for lung nodule diagnosis [15]. We treat the LNDb in the LIDC-IDRI manner, with 451 benign and 768 malignant nodules.

**TianChi:** The TianChi lung disease diagnosis dataset contains a total of 1,470 CT scans with four diseases. There

---

**Table 1.** Comparison of SCL with state-of-the-art SSL methods on the three datasets (\(\text{Mean}(\%) \pm \text{std}(\%)\))

|           | LIDC-IDRI | LNDb   | TianChi |
|-----------|-----------|--------|---------|
|           | AUC       | Accuracy | AUC       | Accuracy | AUC       | Accuracy |
| Context (NIA) [8] | 64.93 ± 1.60 | 56.88 ± 0.97 | 64.57 ± 0.32 | 62.50 ± 0.58 | 64.13 ± 1.05 |
| RotNet (NIA) [9] | 64.96 ± 1.14 | 56.61 ± 2.21 | 63.80 ± 0.59 | 58.40 ± 0.68 | 60.22 ± 0.39 |
| MoCo (NIA) [2] | 71.07 ± 0.11 | 71.29 ± 0.13 | 66.09 ± 0.48 | 67.97 ± 0.46 | 73.44 ± 0.67 |
| MoCo V2 (NIA) [3] | 71.88 ± 0.33 | 71.83 ± 0.25 | 68.20 ± 0.26 | 70.31 ± 0.44 | 76.13 ± 0.72 |
| SimCLR (NIA) [1] | 78.58 ± 0.52 | 79.04 ± 0.50 | 69.71 ± 0.69 | 73.23 ± 0.91 | 79.08 ± 0.56 |
| BYOL (NIA) [10] | 78.35 ± 0.22 | 65.19 ± 0.51 | 63.12 ± 0.33 | 67.97 ± 0.70 | 63.04 ± 0.85 |
| SimSiam (NIA) [11] | 76.39 ± 0.93 | 69.11 ± 1.12 | 68.20 ± 0.44 | 71.88 ± 0.46 | 73.93 ± 0.43 |
| MoCo (MIA) [2] | 73.27 ± 0.55 | 74.15 ± 0.53 | 70.44 ± 0.15 | 70.02 ± 0.21 | 75.63 ± 0.34 |
| MoCo V2 (MIA) [3] | 78.45 ± 0.85 | 78.69 ± 0.55 | 70.30 ± 0.23 | 70.35 ± 0.35 | 77.13 ± 0.65 |
| SimCLR (MIA) [1] | 79.98 ± 0.50 | 80.31 ± 0.71 | 71.21 ± 0.96 | 75.00 ± 1.47 | 80.72 ± 0.93 |
| BYOL (MIA) [10] | 70.40 ± 0.18 | 65.77 ± 0.94 | 69.06 ± 0.26 | 71.87 ± 0.31 | 69.27 ± 0.33 |
| SimSiam (MIA) [11] | 77.10 ± 1.12 | 70.32 ± 1.69 | 72.30 ± 0.36 | 74.72 ± 0.27 | 75.38 ± 0.59 |
| Models Genesis (MIA) [4] | 76.29 ± 2.33 | 65.83 ± 1.56 | 61.73 ± 0.34 | 61.94 ± 0.30 | 63.93 ± 0.43 |
| Rubik’s Cube+ (MIA) [5] | 82.07 ± 0.44 | 81.21 ± 0.16 | 75.63 ± 0.51 | 66.67 ± 0.39 | 76.14 ± 0.73 |
| Restoration (MIA) [12] | 85.60 ± 0.31 | 78.75 ± 0.84 | 74.09 ± 0.51 | 67.71 ± 0.60 | 79.27 ± 0.95 |

SCL (Ours) 93.89 ± 0.60 89.72 ± 0.38 77.95 ± 0.78 82.09 ± 0.67 90.16 ± 0.02
The weight decay rate is 1 with an early-stopping of 50. We set the batch size to 64, and used to optimize the SCL. The training epoch is set to 1000 connected layer. Adam with a base learning rate of 0.001 is used to optimize the SCL.

We transform each 3D lesion to a corresponding 2D view by the 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, 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.

Table 2. Fine-tuning with 10% samples on the three datasets (Mean(%) ± std(%) )

|                     | LIDC-IDRI       | LNDb       | TianChi     |
|---------------------|----------------|------------|-------------|
|                     | AUC             | Accuracy   | AUC         | Accuracy   | Accuracy   |
| Supervised          | 90.90 ± 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 | 77.43 ± 0.85 | 80.25 ± 0.53 | 78.13 ± 0.75 |
| SimCLR (NIA) [1]    | 85.91 ± 0.74    | 84.90 ± 0.73 | 73.62 ± 0.35 | 72.97 ± 0.99 | 79.08 ± 0.56 |
| SimCLR (MIA) [1]    | 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    | 85.75 ± 1.01 | 69.16 ± 1.34 | 78.20 ± 0.63 | 87.85 ± 0.09 |

Fig. 4. Fine-tuning with different numbers of samples.

are 3,264 nodules, 3,613 Streak shadows, 4,201 arteriosclerosis or calcification and 1,140 lymph node calcification.

3.2. Implementation Details

We transform each 3D lesion to a corresponding 2D view by spiral transformation, 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 [16] as base encoder without fully connected layer. Adam 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 using a single Nvidia RTX 2080ti 11GB GPU.

3.3. 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 1. Comparing the existing sota SSL methods, the accuracy of the MIA-based methods is higher than the NIA-based methods. Comparing SCL with other methods, we obtain sota results on all three datasets. The results show that the proposed SCL has superior performance than previous SSL methods.

3.4. 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 2, we fine-tune the SCL with 10% on three datasets, and get the comparable results to Supervised (3D).

As shown in Fig. 4, We also fine-tune the SCL with different percentages of samples. From the overall trend, the accuracy increases with the increase of the samples. The accuracy of SCL also reaches a high value with very few samples (e.g., 1%). It proves that SCL is able to learn high-quality lesion representations.

4. 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 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.
5. REFERENCES

[1] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, “A simple framework for contrastive learning of visual representations,” in International conference on machine learning. PMLR, 2020, pp. 1597–1607.

[2] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross 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.

[3] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He, “Improved baselines with momentum contrastive learning,” arXiv preprint arXiv:2003.04297, 2020.

[4] Zongwei Zhou, Vatsal Sodha, Md Mahfuzur Rahman Siddiquee, Ruibin Feng, Nima Tajbakhsh, Michael B Gotway, and Jianming 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.

[5] Juwen Zhu, Yuexiang Li, Yifan Hu, Kai Ma, S Kevin Zhou, and Yefeng Zheng, “Rubik’s cube+: A self-supervised feature learning framework for 3d medical image analysis,” Medical image analysis, vol. 64, pp. 101746, 2020.

[6] Xiahan Chen, Xiaozhu Lin, Qing Shen, and Xiaohua 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.

[7] Jiahui Wang, Roger Engelmann, and Qiang 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.

[8] Carl Doersch, Abhinav Gupta, and Alexei A Efros, “Unsupervised visual representation learning by context prediction,” in Proceedings of the IEEE international conference on computer vision, 2015, pp. 1422–1430.

[9] Nikos Komodakis and Spyros Gidaris, “Unsupervised representation learning by predicting image rotations,” in International Conference on Learning Representations (ICLR), 2018.

[10] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al., “Bootstrap your own latent: A new approach to self-supervised learning,” arXiv preprint arXiv:2006.07733, 2020.

[11] Xinlei Chen and Kaiming He, “Exploring simple siamese representation learning,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 15750–15758.

[12] Liang Chen, Paul Bentley, Kensaku Mori, Kazunari Misawa, Michitaka Fujiwara, and Daniel Rueckert, “Self-supervised learning for medical image analysis using image context restoration,” Medical image analysis, vol. 58, pp. 101539, 2019.

[13] Samuel G Armato III, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, Binseng Zhao, Denise R Aberle, Claudia I Henschke, Eric 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.

[14] Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Geert Litjens, Paul Gerke, Colin Jacobs, Sarah J Van Riel, Mathilde Marie Winkler Wille, Matiullah Naqibullah, Clara I Sánchez, and Bram 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.

[15] João Pedrosa, Guilherme Aresta, Carlos Ferreira, Márcio Rodrigues, Patrícia Leitão, André Silva Carvalho, João Rebelo, Eduardo Negrão, Isabel Ramos, António Cunha, et al., “Lndb: a lung nodule database on computed tomography,” arXiv preprint arXiv:1911.08434, 2019.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.