MULTIVIEW LONG-SHORT SPATIAL CONTRASTIVE LEARNING FOR 3D MEDICAL IMAGE ANALYSIS

Gongpeng Cao[1], *Yiping Wang*[1], *Manli Zhang*[1], *Jing Zhang*[1], *Guixia Kang*[1,⋆], *Xin Xu*[2]

[1]Key Laboratory of Universal Wireless Communications, Ministry of Education, Beijing University of Posts and Telecommunications, Beijing 100876, China
[2]Department of Neurosurgery, General Hospital of PLA, Beijing 100853, China

ABSTRACT

The success of supervised deep learning heavily depends on large labeled datasets whose construction is often challenging in medical image analysis. Contrastive learning, a variant of self-supervised learning, is a potential solution to alleviate the strong demand for data annotation. In this work, we extend the contrastive learning framework to 3D volumetric medical imaging. Specifically, we propose (1) multiview contrasting strategy to maximize the mutual information between three views of 3D image to learn global representations and (2) long-short spatial contrasting strategy to learn local representations by matching a short spatial clip to a long spatial clip in the latent space. To combine these two strategies, we propose multiview long-short spatial contrastive learning (MLSSCL) framework, which can effectively learn generic 3D representations. Our extensive experiments on two brain Magnetic Resonance Imaging (MRI) datasets demonstrate that MLSSCL significantly outperforms learning from scratch and other self-supervised learning methods on both classification and segmentation tasks.

Index Terms— Contrastive learning, 3D representation learning, multiview contrast, long-short spatial contrast, medical image analysis

1. INTRODUCTION

Deep convolutional neural networks (CNNs) have achieved significant success on many challenging tasks in medical image analysis such as disease classification [1, 2] and lesion segmentation [3, 4, 5]. A key requirement for many of these successes is a sufficient number of labeled training data. However, annotating volumetric medical images is not only time-consuming and laborious, but it also requires expert knowledge in specific fields, which makes it difficult to obtain a large labeled dataset. As a result, although various 3D network architectures have been proposed, their performance is often confined by the lack of expert annotation.

To deal with the deficiency of annotated data, researchers have attempted to use self-supervised learning (SSL) methods to mine useful information from the unlabeled raw data, which can offer a great initialization for downstream tasks with limited labeled data. Various pretext task-based self-supervised methods have been proposed for 3D medical imaging. Examples include context restoration [6], rubik's cube recovery [7] and restoring the deformed sub-volumes [8]. Although these methods have achieved good results, heuristic pretext tasks tend to focus on partial information in the data, resulting in biased representations and limited generalization. Self-supervised contrastive learning [9, 10, 11, 12] have shown great promise to learn generic representations in computer vision field over the last year. The idea behind this approach is to directly learn instance-level representations by pulling similar samples together in the latent space while pushing dissimilar samples apart.

However, there are two limitations when applying contrastive learning methods to 3D medical image analysis. Firstly, the contrasting strategy is often devised based on data augmentation, and does not utilize the intrinsic structural similarity that is potentially present in 3D medical images. Secondly, most methods just extract global representations and ignore the local representations. To address these issues, [13] extended the contrasting strategy by leveraging real valued proxy metadata for classification of volumetric medical images and [14] extended the contrastive learning framework for segmentation of medical images by leveraging domain-specific and problem-specific cues with 2D CNNs.

In this work, we focus in devising contrasting strategies by leveraging the multiview and long-short spatial information of 3D volumetric medical images and make the following contributions. (1) We introduce multiview contrasting strategy to learn global representations by maximizing the mutual information between three views of the same volumetric medical image. (2) We introduce long-short spatial contrasting strategy to learn local representations by matching a short spatial clip to a long spatial clip in the latent space under

⋆Corresponding author: gxkang@bupt.edu.cn

This work was supported by Fundamental Research Funds for the Central Universities (2020XD-A06-1), the State Key Program of the National Natural Science Foundation of China (82030037), the National Science and Technology Major Project of China (No.2017ZX03001022).

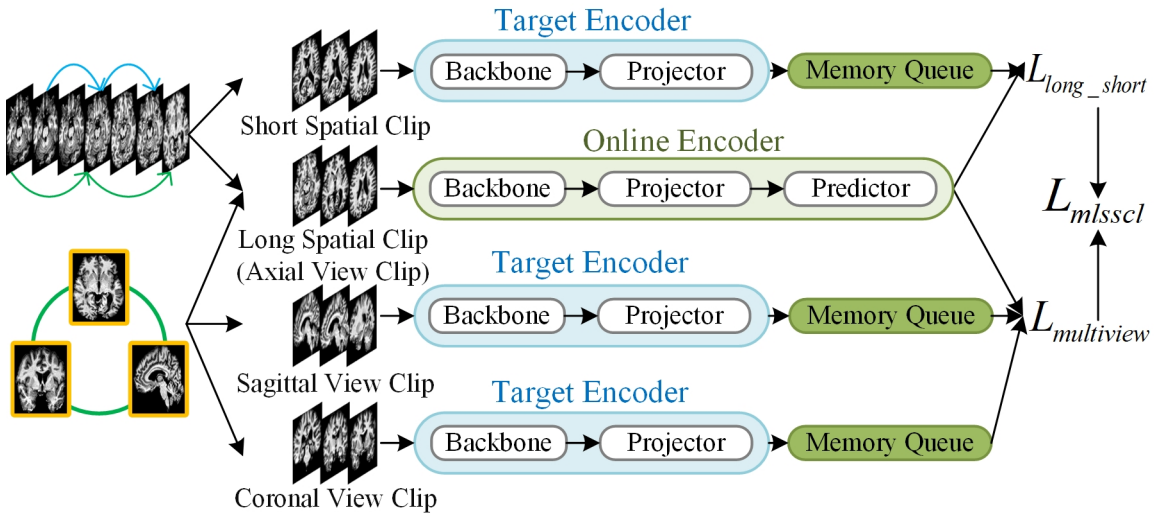

Fig. 1. Multiview Long-Short Spatial Contrastive Learning Framework. The framework involves two contrasting strategies: multiview contrasting strategy enables model to learn global representations and long-short spatial contrasting strategy enables model to learn local representations.

the given view. (3) We propose multiview long-short spatial contrastive learning (MLSSCL) framework to combine these two contrasting strategies, which can effectively learn generic 3D representations. (4) We evaluate MLSSCL on two brain Magnetic Resonance Imaging (MRI) datasets and the experimental results show that our method achieves better results compared to learning from scratch and other state-of-the-art (SOTA) SSL methods.

2. METHOD

2.1. Contrastive Learning Review

Given two random variables v_1 and v_2, contrastive learning aims to learn a parametric function $h(\cdot,\cdot)$ to discriminate between "positive" samples drawn from the empirical joint distribution $p(v_1)p(v_2|v_1)$ and "negative" samples drawn from the product of marginals $p(v_1)p(v_2)$ based on InfoNCE loss [15]. In practice, for a given anchor point v_1^i, the InfoNCE loss L_N optimizes $h(\cdot,\cdot)$ to achieve a higher value for the correct positive $v_2^i \sim p(v_2|v_1^i)$ compared to a set of K distractors $v_2^j \sim p(v_2)$:

$$L_N(v_1, v_2) = -\mathbb{E}\left[log\frac{e^{h(v_1^i, v_2^i)}}{\sum_{j=1}^{K+1} e^{h(v_1^i, v_2^j)}}\right] \quad (1)$$

2.2. Multiview Contrasting Strategy

Inspired by [16], we assume that the information shared between three views can capture the global representations of volumetric medical image. This corresponds to the fact that physicians usually need to observe all three views of 3D medical image to get comprehensive information.

Therefore, to learn the global representations, we need to maximize the mutual information between three views of volumetric image. Formally, denoting the axial view, coronal view and sagittal view of 3D medical image as v_a, v_c and v_s respectively, we formulate the problem as:

$$\max\{I(v_a; v_c) + I(v_a; v_s) + I(v_c; v_s)\} \quad (2)$$

where $I(\cdot;\cdot)$ denotes mutual information. However, it is notoriously difficult to compute the mutual information for real valued high-dimensional data. Fortunately, [15] has proved that InfoNCE loss can estimate the lower bound of mutual information, i.e., for two views v_1, v_2:

$$I(v_1; v_2) \geq log(K) - L_N(v_1, v_2) \quad (3)$$

Given any K, minimizing $L_N(v_1, v_2)$ maximizes the lower bound on the mutual information $I(v_1; v_2)$.

Therefore, we transform the problem of maximizing mutual information between three views into a multiview contrastive learning problem. According to (2) and (3), we propose multiview contrastive loss:

$$L_{multiview} = L_N(v_a, v_c) + L_N(v_a, v_s) + L_N(v_c, v_s) \quad (4)$$

As presented in Fig. 1, by minimizing $L_{multiview}$, we maximize the mutual information between three views for each sample.

2.3. Long-Short Spatial Contrasting Strategy

The goal of our long-short spatial contrasting strategy is to to learn local representations by maximizing representation similarity between a long spatial clip v_L and a much shorter spatial clip v_S where both clips are sampled from the same 3D

volumetric medical image under the given view. To achieve this goal, we propose long-short spatial contrastive loss:

$$L_{long\_short} = L_N(v_S, v_L) \tag{5}$$

As presented in Fig. 1, by matching the short-clip representation to the long-clip representation, the model can extrapolate the contextual information exhibited in the long spatial clip, which is beneficial to understanding and recognizing the structure and correlation of local tissues in volumetric medical images.

2.4. Multiview Long-Short Spatial Contrastive Learning Framework

To effectively learn generic 3D representations, we propose MLSSCL to integrate above two contrasting strategies. Below we describe specific details related to MLSSCL.

Network Architecture. As illustrated in Fig. 1, MLSSCL is comprised of four encoders: an online encoder and three target encoders which have shared weights and will be discarded after pre-training. There are a backbone and a projector head (2-layer multi-layer perceptron (MLP)) in all four encoders. A prediction head (2-layer MLP) is added to the online encoder to introduce asymmetry. Following [12], the online encoder is updated by back-propagation and the target encoders are updated in the manner of momentum to keep the representations' consistency. Besides, every target encoder has a memory queue, which is updated in each training iteration, to store previous representations.

Clip Sampling and Contrasting Strategy. Considering a batch B of unlabeled training 3D volumetric medical images, we randomly sample an axial view clip, a coronal view clip and a sagittal view clip from each of them. Each clip includes C slices and has the same sampling spatial stride δ_L. To work with long-short spatial contrasting strategy, we regard the above axial view clip as the long spatial clip and randomly sample C axial slices with spatial stride δ_S ($\delta_S < \delta_L$) as a short spatial clip for each sample. We denote the five clip sets as $V_L = V_a = \{v_a^i\}_{i=1}^B = \{v_L^i\}_{i=1}^B, V_c = \{v_c^i\}_{i=1}^B, V_s = \{v_s^i\}_{i=1}^B, V_S = \{v_S^i\}_{i=1}^B$ repectively, as shown in Fig. 1. Based on (4) and (5), we get contrastive loss for MLSSCL:

$$L_{mlsscl} = \alpha L_{multiview} + \beta L_{long\_short} \tag{6}$$

where α and β are two hyperparameters to adjust the weights of two strategies. $L_{multiview}$ trains the model to learn global representations and L_{long_short} trains the model to learn local representations. The combination of two strategies enables the model to learn universal 3D representations.

3. EXPERIMENTS

3.1. Experimental Setup

Datasets

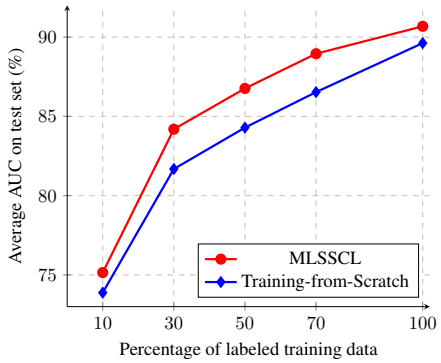

Fig. 2. The AD classification performance of networks trained with different amounts of labeled data.

- *Alzheimer's Disease Neuroimaging Initiative (ADNI) Database.* We collected 7461 T1-weighted MRI scans from ADNI database to build a large-scale brain MRI dataset. Subjects are randomly assigned into ADNI pre-training set (5953 scans) and ADNI test set (1538 scans). Besides, we randomly sample 332 healthy controls (HC), 332 Alzheimer's disease (AD) from the ADNI pre-training set as ADNI-AD classification training set and 82 HC, 82 AD from the ADNI test set as ADNI-AD classification test set. All images have been pre-processed as in [1].

- *ISBI 2015 Longitudinal Multiple Sclerosis (MS) Lesion Segmentation Challenge Dataset.* This dataset [17] consists of 5 patients for training and 14 patients for test. Each patient has 4 or 5 time-points and each time-point includes four kinds of MR sequences: T1, T2, PD and FLAIR. The labels for segmentation include the background (label 0) and white matter lesion (label 1).

Implementation Details

- *Instantiation of Network.* For AD classification task, we use the modified 3D ResNet-18 as backbone, where we modify the kernel size of the first convolution layer to 3x3x3. For MS lesion segmentation task, we use the 3D UNet-based encoder as backbone. All backbones do not downsample in the direction of clip sampling during pre-training. The projector head and prediction head keep the same with [12], except that the dimensions of input layer and output layer are set to (1024, 128) for 3D ResNet-18 and (512, 128) for 3D UNet-based encoder. The memory queue length and starting momentum of the target networks are set to 320 and 0.99.

- *Instantiation of Contrasting Strategy.* We set $B = 64, C = 16, \delta_S = 2, \delta_L = 4, \alpha = 0.5$ and $\beta = 0.5$. Besides, we use the following data augmentation techniques: random cropping and resizing, mirror flip, intensity scaling, intensity shift and gaussian blur.

- *Optimization Details.* We pre-train models on ADNI pre-training set for 100 epochs. SGD optimizer is used for optimization with initial learning rate of 0.03 and a cosine decay learning rate schedule. For AD classification and MS lesion segmentation, we keep the same with [8] and [5] respectively.

Table 1. Results(mean±std) for AD classification (AD vs. HC) on the ADNI-AD classification test set.

Method	ACC	SEN	SPE	AUC
Training-from-Scratch	0.793 ± 0.011	0.874 ± 0.055	0.711 ± 0.058	0.896 ± 0.006
BYOL [12]	0.809 ± 0.004	0.866 ± 0.032	0.752 ± 0.037	0.886 ± 0.016
MoCo [10]	0.825 ± 0.020	0.886 ± 0.043	0.764 ± 0.060	0.895 ± 0.001
Model Genesis [8]	0.827 ± 0.004	0.911 ± 0.061	0.744 ± 0.061	0.904 ± 0.009
Age-Aware [13]	0.831 ± 0.007	0.882 ± 0.007	0.780 ± 0.012	0.899 ± 0.011
MLSSCL	$\mathbf{0.858 \pm 0.013}$	$\mathbf{0.911 \pm 0.019}$	$\mathbf{0.805 \pm 0.044}$	$\mathbf{0.907 \pm 0.012}$

Table 2. The segmentation results of different approaches on the ISBI 2015 longitudinal MS lesion segmentation test set.

Method	DSC†	PPV†	LTPR†	LFPR†
Training-from-Scratch	0.6176	0.8229	0.4451	0.3485
SSL				
Age-Aware [13]	0.6320	0.8103	0.4586	0.3034
BYOL [12]	0.6337	0.7991	0.4675	0.3442
MoCo [10]	0.6369	0.7972	0.4641	0.3092
Model Genesis [8]	0.6434	0.8200	0.4647	0.3082
MS SOTA				
Aslania et al. [3]	0.6114	**0.8992**	0.4103	0.1393
Andermatt et al. [18]	0.6298	0.8446	0.4870	0.2013
Valverde et al. [4]	0.6304	0.7866	0.3669	0.1529
Hu et al. [5]	0.6345	0.8682	0.4787	**0.1299**
MLSSCL	**0.6482**	0.8007	**0.4933**	0.2796

† Detailed definitions of these evaluation metrics can be found in [17].

3.2. Evaluation on AD Classification

We first fine-tune the pre-trained networks for AD classification to evaluate the learned representations. Specifically, we adopt 3-fold cross validation on ADNI-AD classification training set and report the average accuracy(ACC), sensitivity (SEN), specificity (SPE) and area under curve (AUC) on ADNI-AD classification test set. As shown in Table 1, MLSSCL achieves a remarkable improvement over training-from-scratch, increasing the ACC by 6.5%, SEN by 3.7% and SPE by 9.4%. This result confirms that MLSSCL can provide better initialization for the target task model. Meanwhile, compared to other SSL methods, MLSSCL also yields at least 2.7% improvement for ACC and 2.5% for SPE.

To better demonstrate the transferability of MLSSCL, we fine-tune the pre-trained network with different amounts of labeled data (i.e., 10%, 30%, 50% and 70%). As presented in Fig. 2, MLSSCL can effectively deal with the situation with few labeled training samples. Specifically, using 70% labeled data, MLSSCL can achieve an AUC of 88.95% similar to training-from-scratch with 100% training data.

3.3. Evaluation on MS Lesion Segmentation

To comprehensively evaluate our framework, the pre-trained networks are transferred to MS lesion segmentation task. Following [2], we duplicate the weights of pre-trained input layer by the number of modalities to process multimodal inputs. The results on the test set assessed by the online evaluation server are presented in Table 2. It can be observed that MLSSCL consistently outperforms training-from-scratch and other SSL methods. Specifically, MLSSCL produces a gain of 3.06% for DSC, 4.82% for LTPR and 6.89% for LFPR in comparison with training-from-scratch. This implies that MLSSCL can not only effectively learn global representations, but also capture the structure of sophisticated local tissues. Besides, we also compare with MS lesion segmentation SOTA methods which use more complex models. As shown in Table 2, MLSSCL still achieves higher DSC and LTPR, demonstrating that MLSSCL's generic and robust representations can greatly alleviate over-fitting problem.

Table 3. Ablation to contrasting strategies on AD classification task (mean ± std).

Contrasting Strategy	ACC	AUC
Long-Short	0.823 ± 0.012	0.892 ± 0.014
Multiview	0.833 ± 0.027	0.906 ± 0.024
Multiview & Long-Short	$\mathbf{0.858 \pm 0.013}$	$\mathbf{0.907 \pm 0.012}$

3.4. Ablation to Contrasting Strategies

To investigate the effect of proposed contrasting strategies, we pre-train models with multiview contrasting strategy, long-short spatial contrasting strategy and both contrasting strategies (MLSSCL) respectively. As presented in Table 3, the performance of using both contrasting strategies is better than just using only one, which indicates the importance of effective combination of global representations and local representations.

4. CONCLUSION

In this paper, we proposed multiview long-short spatial contrastive learning (MLSSCL) framework for 3D visual representations learning with unlabeled volumetric medical images, involving two contrasting strategies–multiview contrasting strategy and long-short spatial contrasting strategy. The former contrasting strategy forced the 3D neural networks to learn global representations, while the latter one improved the ability of models to capture local representations. The proposed MLSSCL was evaluated on AD classification task and MS lesion segmentation task. The experimental results showed that MLSSCL remarkablely outperformed training-from-scratch method, especially when fine-tuned on only small amounts of labeled data, and also showed a clear superiority compared with other self-supervised learning methods.

5. REFERENCES

[1] Wenyong Zhu, Liang Sun, Jiashuang Huang, Liangxiu Han, and Daoqiang Zhang, "Dual attention multi-instance deep learning for alzheimer's disease diagnosis with structural mri," *IEEE Transactions on Medical Imaging*, 2021.

[2] Aiham Taleb, Winfried Loetzsch, Noel Danz, Julius Severin, Thomas Gaertner, Benjamin Bergner, and Christoph Lippert, "3d self-supervised methods for medical imaging," *arXiv preprint arXiv:2006.03829*, 2020.

[3] Shahab Aslani, Michael Dayan, Loredana Storelli, Massimo Filippi, Vittorio Murino, Maria A Rocca, and Diego Sona, "Multi-branch convolutional neural network for multiple sclerosis lesion segmentation," *NeuroImage*, vol. 196, pp. 1–15, 2019.

[4] Sergi Valverde, Mariano Cabezas, Eloy Roura, Sandra González-Villà, Deborah Pareto, Joan C Vilanova, Lluís Ramió-Torrentà, Àlex Rovira, Arnau Oliver, and Xavier Lladó, "Improving automated multiple sclerosis lesion segmentation with a cascaded 3d convolutional neural network approach," *NeuroImage*, vol. 155, pp. 159–168, 2017.

[5] Chuan hu, Guixia Kang, Beibei Hou, Yiyuan Ma, Fabrice Labeau, and Zichen Su, "Acu-net: A 3d attention context u-net for multiple sclerosis lesion segmentation," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 1384–1388.

[6] 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.

[7] Xinrui Zhuang, Yuexiang Li, Yifan Hu, Kai Ma, Yujiu Yang, and Yefeng Zheng, "Self-supervised feature learning for 3d medical images by playing a rubik's cube," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 420–428.

[8] Zongwei Zhou, Vatsal Sodha, Jiaxuan Pang, Michael B Gotway, and Jianming Liang, "Models genesis," *Medical image analysis*, vol. 67, pp. 101840, 2021.

[9] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," *arXiv preprint arXiv:2006.09882*, 2020.

[10] 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.

[11] 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.

[12] 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.

[13] Benoit Dufumier, Pietro Gori, Julie Victor, Antoine Grigis, Michel Wessa, Paolo Brambilla, Pauline Favre, Mircea Polosan, Colm McDonald, Camille Marie Piguet, et al., "Contrastive learning with continuous proxy meta-data for 3d mri classification," *arXiv preprint arXiv:2106.08808*, 2021.

[14] Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu, "Contrastive learning of global and local features for medical image segmentation with limited annotations," *arXiv preprint arXiv:2006.10511*, 2020.

[15] Aaron van den Oord, Yazhe Li, and Oriol Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[16] Yonglong Tian, Dilip Krishnan, and Phillip 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.

[17] Aaron Carass, Snehashis Roy, Amod Jog, Jennifer L Cuzzocreo, Elizabeth Magrath, Adrian Gherman, Julia Button, James Nguyen, Ferran Prados, Carole H Sudre, et al., "Longitudinal multiple sclerosis lesion segmentation: resource and challenge," *NeuroImage*, vol. 148, pp. 77–102, 2017.

[18] Simon Andermatt, Simon Pezold, and Philippe C Cattin, "Automated segmentation of multiple sclerosis lesions using multi-dimensional gated recurrent units," in *International MICCAI Brainlesion Workshop*. Springer, 2017, pp. 31–42.