

Thyroid ultrasound diagnosis improvement via multi-view self-supervised learning and two-stage pre-training

Jian Wang^a, Xin Yang^{b,c,d}, Xiaohong Jia^e, Wufeng Xue^{b,c,d}, Rusi Chen^{b,c,d}, Yanlin Chen^{b,c,d}, Xiliang Zhu^{b,c,d}, Lian Liu^{b,c,d}, Yan Cao^f, Jianqiao Zhou^{e,*}, Dong Ni^{b,c,d,*} and Ning Gu^{a,g,*}

^aKey Laboratory for Bio-Electromagnetic Environment and Advanced Medical Theranostics, School of Biomedical Engineering and Informatics, Nanjing Medical University, Nanjing, 211166, China

^bNational-Regional Key Technology Engineering Laboratory for Medical Ultrasound, School of Biomedical Engineering, Health Science Center, Shenzhen University, Shenzhen, 518073, China

^cMarshall Laboratory of Biomedical Engineering, Shenzhen University, Shenzhen, 518073, China

^dMedical UltraSound Image Computing (MUSIC) Lab, Shenzhen University, Shenzhen, 518073, China

^eDepartment of Ultrasound, Ruijin Hospital, Shanghai Jiaotong University School of Medicine, Shanghai, 200025, China

^fShenzhen RayShape Medical Technology Co., Ltd, Shenzhen, 518051, China

^gCardiovascular Disease Research Center, Nanjing Drum Tower Hospital, Affiliated Hospital of Medical School, Medical School, Nanjing University, Nanjing, 210093, China

ARTICLE INFO

Keywords:

Thyroid ultrasound image
Self-supervised learning
Multi-view learning
Two-stage pre-training
Nodule classification
Nodule segmentation

ABSTRACT

Thyroid nodule classification and segmentation in ultrasound images are crucial for computer-aided diagnosis; however, they face limitations owing to insufficient labeled data. In this study, we proposed a multi-view contrastive self-supervised method to improve thyroid nodule classification and segmentation performance with limited manual labels. Our method aligns the transverse and longitudinal views of the same nodule, thereby enabling the model to focus more on the nodule area. We designed an adaptive loss function that eliminates the limitations of the paired data. Additionally, we adopted a two-stage pre-training to exploit the pre-training on ImageNet and thyroid ultrasound images. Extensive experiments were conducted on a large-scale dataset collected from multiple centers. The results showed that the proposed method significantly improves nodule classification and segmentation performance with limited manual labels and outperforms state-of-the-art self-supervised methods. The two-stage pre-training also significantly exceeded ImageNet pre-training.

1. Introduction

Thyroid cancer is among the most common cancers worldwide (Sung, Ferlay, Siegel, Laversanne, Soerjomataram, Jemal and Bray, 2021). Early detection enables timely intervention and avoids overdiagnosis (Bethesda, 2018). Ultrasound is the primary imaging tool for thyroid diagnosis because it is real-time, non-invasive, and low-cost (Smith-Bindman, Miglioretti, Johnson, Lee, Feigelson, Flynn, Greenlee, Kruger, Hornbrook, Roblin et al., 2012). However, accurate interpretation of ultrasound images requires experienced physicians, and inexperienced physicians may misdiagnose. Several computer-aided diagnostic methods have been proposed to address this issue, most of which are based on deep learning. For example, Deng, Han, Wei and Chang (2022) proposed a multi-task network to determine Thyroid Imaging Reporting and Data System grade for identifying the benignity and malignancy of thyroid nodules. Sun, Wu, Zhao, Gao, Xie, Lin, Sui, Li, Wu and Ni (2023) proposed a contrast-learning-based thyroid nodule classification model to improve the accuracy of diagnosis. Kang, Lao, Li, Jiang, Qiu, Zhang and Li (2022) proposed intra- and inter-task consistent learning to enforce the network to learn consistent predictions for

nodule classification and segmentation. Gong, Chen, Chen, Li, Li and Chen (2023) designed a multi-task learning framework to accurately segment the thyroid nodule. For more related work, please refer to Chen, You and Li (2020a); Sharifi, Bakhshali, Dehghani, DanaiAshgzari, Sargolzaei and Eslami (2021).

Although these methods have the potential to solve the aforementioned clinical problem, they require large amounts of data with manual annotations. The fine-needle aspiration (FNA) biopsy is the gold standard for distinguishing benign and malignant nodules. However, only a small fraction of patients undergo an FNA biopsy. For nodule segmentation, the gold standard is obtained by manually delineating the pixel-level mask. As shown in Fig.1, the borders of the nodules in the image are often blurred and incomplete and the nodules may resemble carotid vessels. Therefore, annotating the mask requires a comprehension of the thyroid anatomy, recognition of relevant features from the ultrasound image, and exclusion of interference from similar tissues. Annotating ultrasound images is time-consuming and requires costly expertise that is not easily accessible. Most labeled thyroid image datasets are thus small and lack manual labels. This limits the development of deep learning techniques in thyroid image analysis.

Recently, self-supervised learning (SSL) has been applied to various medical images to address the problem of insufficient labeled data. Its core objective is to provide good initialization for the target task by pre-training the model

*Corresponding author

✉ zhousu30@126.com (J. Zhou); nidong@szu.edu.cn (D. Ni);

guning@nju.edu.cn (N. Gu)

ORCID(s):

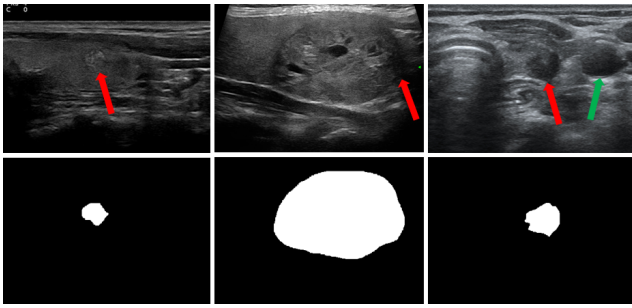


Figure 1: The upper row is the thyroid ultrasound images of different patients, and the lower row is the corresponding nodule masks. Red arrows indicate nodules and green arrows indicate carotid vessels.

without using artificial labels. For example, Zhou, Sodha, Pang, Gotway and Liang (2021c) used image restoration to pre-train models on 2D chest CT slices, chest X-ray images, and 3D chest volumes. Zhou, Yu, Bian, Hu, Ma and Zheng (2020) used contrastive learning to pre-train the model on chest X-ray images. Zhu, Li, Hu, Ma, Zhou and Zheng (2020) used 3D jigsaw puzzles to pre-train the model on brain CT and MRI volumes. Punn and Agarwal (2022) pre-trained models on breast ultrasound images and dermoscopy images via redundancy reduction. Basu, Singla, Gupta, Rana, Gupta and Arora (2022) combined contrastive learning and hard negative mining to pre-train models on gallbladder ultrasound videos. For more information about the medical imaging applications of SSL, please refer to Shurrah and Duwairi (2022).

We investigated a large number of SSL-related research in the field of medical image analysis from the literature and found that the existing methods have some limitations. *First*, most methods were designed for single-view tasks, which does not consider the multi-view nature of the thyroid. Physicians usually scan the thyroid gland of a patient horizontally and vertically to conduct a complementary examination, resulting in transverse and longitudinal views. It is also possible that one of the two views is missing. As shown in Fig.2, the lower row shows thyroid transverse views from four patients, and the upper row shows the corresponding longitudinal views. Two views of the same nodule display relevant and complementary information, such as nodule shape and echo pattern. Moreover, two views of the same nodule should belong to the same category, either benign or malignant. Intuitively, exploiting such multi-view consistency can enhance SSL, thereby improving the target task performance. *Second*, the SSL methods designed for multimodal data can theoretically be used for multi-view thyroid images (Hervella, Rouco, Novo and Ortega, 2020, 2021; Fedorov, Sylvain, Geenjaer, Luck, Wu, DeRamus, Kirilin, Bleklov, Calhoun and Plis, 2021a; Fedorov, Wu, Sylvain, Luck, DeRamus, Bleklov, Plis and Calhoun, 2021b; Li, Jia, Islam, Yu and Xing, 2020; Taleb, Lippert, Klein and Nabi, 2017; Xiang, Zhuo, Zhao, Deng, Zhu, Wang, Jiang and Lei, 2022), but they cannot handle missing views well. This is because they all require the paired data, and

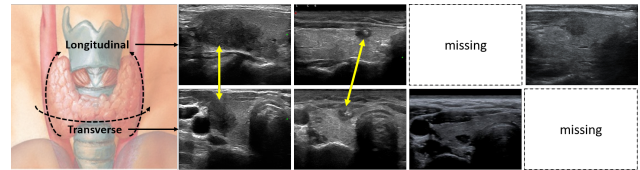


Figure 2: The color image on the left is a schematic of the thyroid, and thyroid ultrasound images from four patients are on the right. The upper row is the longitudinal views, and the lower row is the corresponding transverse views. The yellow arrows point to the same nodule.

unpaired data cannot be used. *Third*, the combination of SSL-based pre-training on medical images and pre-training on natural images may benefit the target task but was ignored by existing self-supervised studies. Most self-supervised studies have demonstrated that self-supervised pre-training on medical images can outperform supervised pre-training on natural images for target tasks. However, one can easily collect numerous natural images from the Internet, but it is difficult to obtain numerous medical images from hospitals, even without labels.

In this study, we attempt to improve the nodule classification and segmentation performance of thyroid ultrasound images with limited manual labels using a novel SSL method. The main contributions of our study are as follows: (1) We proposed a multi-view SSL method for thyroid ultrasound image analysis. To address the issue of missing views, we designed an adaptive loss function that allowed the model to utilize unpaired views. (2) We adopted a two-stage pre-training strategy to combine ImageNet pre-training and self-supervised pre-training on thyroid ultrasound images, which exploits the benefits of both. (3) Extensive experiments were conducted on a large-scale thyroid ultrasound image dataset collected from multiple centers and devices. The experimental results showed that the proposed method significantly improves nodule classification and segmentation performance with limited manual labels, demonstrating its effectiveness.

2. Related work

In this section, we briefly introduce three fields related to our research: self-supervised learning, multi-view learning, and two-stage pre-training.

2.1. Self-supervised learning

SSL can be classified into three main categories: predictive, generative, and contrastive (Shurrah and Duwairi, 2022). *Predictive* methods usually first transform the image and then use models to predict these transformations. Examples include the jigsaw puzzle (Noroozi and Favaro, 2016), relative-position prediction (Doersch, Gupta and Efros, 2015), and rotation prediction (Gidaris, Singh and Komodakis, 2018). However, these methods must be carefully designed to prevent the network from taking shortcuts to learn meaningless representations for the target tasks. *Generative* methods usually use an encoder-decoder structure

for image reconstruction, such as image inpainting (Pathak, Krahenbuhl, Donahue, Darrell and Efros, 2016), image context restoration (Chen, Bentley, Mori, Misawa, Fujiwara and Rueckert, 2019), models genesis (Zhou et al., 2021c). Such methods often provide limited improvements in the target performance. Recently, breakthroughs have been achieved in visual transformers (ViTs) based on generative methods (He, Chen, Xie, Li, Dollár and Girshick, 2022). Partial patches of the image were masked, and visible patches were sent to a ViT-based network for patch reconstruction. Although these methods produce more generalizable visual features, they are specifically designed for ViT architecture and are computationally intensive. Additionally, without further fine-tuning, pre-trained features may not perform favorably in some scenarios. *Contrastive* methods aim to minimize the feature distances of positive pairs while maximizing the feature distances between negative pairs. Positive pairs usually refer to different augmented versions of the same image, and negative pairs usually refer to different images. The representative methods include SimCLR (Chen, Kornblith, Norouzi and Hinton, 2020b) and MoCo (He, Fan, Wu, Xie and Girshick, 2020; Chen, Fan, Girshick and He, 2020c). These methods have been shown to outperform supervised ImageNet pre-training. In this study, we extend the common contrastive approach by considering different thyroid ultrasound views of the same nodule as positive pairs.

2.2. Multi-view learning

Multi-view learning is a scenario where representations are learned by correlating information from multiple views of data to improve learning performance (Li, Yang and Zhang, 2018). *Multi-view supervised learning is an active research area* (Wang, Miao, Yang, Li, Zhou, Huang, Lin, Xue, Jia, Zhou et al., 2020; Wu, Xie, Zhu, Ao, Chen, Zhang, Zhuang, Lin and He, 2022; Shah, Shah, Lau, de Melo and Chellappa, 2023; Kim and Song, 2022). For example, Wang et al. (2020) proposed a multimodal fusion network that fuses B-mode, Doppler, shear wave, and strain wave ultrasound images to predict benign and malignant breast nodules. In *multi-view self-supervised learning*, Hervella et al. (2020, 2021) proposed a generative framework that reconstructs gray angiography from the corresponding colorful retinography. Fedorov et al. (2021a,b) used a contrastive framework on multimodal MRI images to maximize the mutual information. Xiang et al. (2022) proposed a self-supervised multi-modal fusion network on multimodal thyroid images. Hassani and Khasahmadi (2020) proposed a contrastive framework on graphs. Roy and Etemad (2021) proposed a contrastive framework on human facial images. However, these SSL methods require paired views. To address this problem, Li et al. (2020) first used CycleGAN (Zhu, Park, Isola and Efros, 2017) to synthesize missing modalities from other modalities. They then pre-trained the model using a contrastive self-supervised framework that aligned multimodal features. Similarly, Taleb et al. (2017) also used CycleGAN to synthesize the missing modality and

designed multimodal jigsaw puzzles on multimodal MRI images. However, the quality of synthetic data is difficult to evaluate and may be detrimental to pre-training. In contrast, the proposed method does not require paired data. Additionally, our method is evaluated on image classification, segmentation, and multi-view classification, which has not been verified in other studies.

2.3. Two-stage pre-training

Given the significant differences between natural and medical images, transfer learning from natural to medical images may be suboptimal (Raghu, Zhang, Kleinberg and Bengio, 2019). Therefore, some studies have proposed a two-stage pre-training for target tasks. For example, Liu, Dong, Wang, Cui, Fan, Ma and Chen (2021) first initialized the partial layers of their proposed network with weights pre-trained on ImageNet and continued to pre-train the model on numerous labeled pulmonary nodule CT images before fine-tuning it for COVID-19 lung infection segmentation. Similarly, Meng, Tan, Yu, Wang and Liu (2022) used two-stage pre-training to initialize the model and used it for COVID-19 image classification. Zhang, Chen, Gao, Huang, Li and Zhang (2022) built a large dataset from natural images similar to tongue manifestation images and trained the pre-trained model on it before fine-tuning it on real clinical tongue manifestation images for target tasks. Although these studies demonstrated the effectiveness of two-stage pre-training, it is challenging to collect a large amount of labeled data for supervised learning in the second stage. A natural idea is to replace supervised pre-training with self-supervised pre-training in the second stage, which exploits a large amount of unlabeled data. However, only a few studies (Azizi, Mustafa, Ryan, Beaver, Freyberg, Deaton, Loh, Karthikesalingam, Kornblith, Chen et al., 2021; Verma and Tapaswi, 2022) have attempted this strategy, and a detailed analysis is lacking. In this study, we carefully explore two-stage pre-training in two ways: supervised to self-supervised and self-supervised to self-supervised.

3. Method

In this section, we first introduce the proposed SSL framework and then propose our adaptive loss. Finally, we introduce the two-stage pre-training and implementation details.

3.1. Proposed framework

Common contrastive learning methods regard different data-augmented versions of the same image as positive pairs, and different images as negative pairs. They reduce the feature distance between positive pairs and increase the feature distance between negative pairs. This allows the model to be trained without manual labels and to achieve a target task performance that meets or exceeds the level of supervised ImageNet pre-training (Chen et al., 2020b,c). However, such methods may lead to multi-view images of the same nodule being assigned to different categories. In thyroid ultrasound scanning, the transverse and longitudinal views of the same

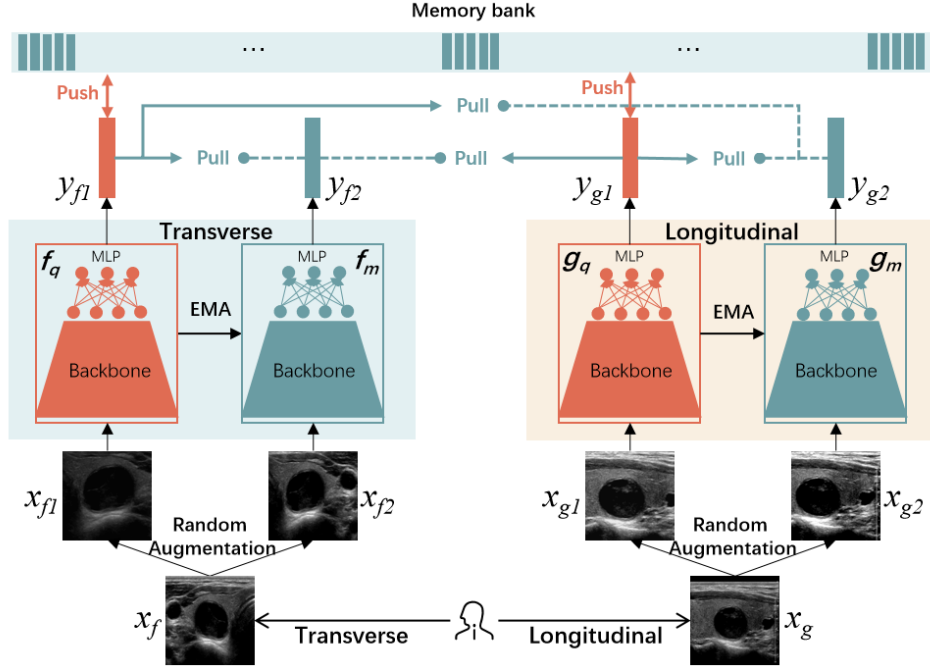


Figure 3: Our framework adopts independent query and momentum encoders for each view, and the two views share the same memory bank.

nodule are visually related. As shown in Fig.2, two views of the same nodule are related and complementary in terms of nodule shape and echo pattern and are differentiated in views of different nodules. Furthermore, two views of the same nodule share the same benign and malignant categories. Therefore, common contrastive learning methods may not be suitable for multi-view thyroid ultrasound images.

To solve this problem, we propose a multi-view contrastive self-supervised framework. Fig.3 shows this framework. Inspired by MoCo v2 (Chen et al., 2020c), we adopt two sets of encoders for the transverse and longitudinal views, where each set of encoders consists of a query and a momentum encoder. For convenience, we denote the transverse and longitudinal encoders as $f(\cdot)$ and $g(\cdot)$, and use the subscripts q and m to represent the query and momentum encoder, respectively. The four encoders ($f_q(\cdot)$, $f_m(\cdot)$, $g_q(\cdot)$, and $g_m(\cdot)$) share the same network architecture and initial weights. Each encoder consists of a backbone and a projection head. The backbone and projection are convolutional neural networks (CNNs) and multilayer perceptron (MLP), respectively. The backbone is used to extract features from the images in the input space, and the projection head is used to project the extracted features into the latent space. After pre-training, only the backbone of the query encoder is used for the target tasks, whereas the projection head is discarded. Additionally, we adopt a memory bank that can store K vectors. This memory bank stores the vectors projected onto the latent space and works as a queue that follows the first-in-first-out principle.

Given the paired views of the same patient from a batch \mathbf{X} , we denote the transverse view as \mathbf{x}_f and the longitudinal view as \mathbf{x}_g . We employ the random data augmentation on \mathbf{x}_f

and \mathbf{x}_g to generate two augmented versions (\mathbf{x}_{f1} and \mathbf{x}_{f2} , \mathbf{x}_{g1} and \mathbf{x}_{g2}). The augmented transverse views, \mathbf{x}_{f1} and \mathbf{x}_{f2} are fed into $f_q(\cdot)$ and $f_m(\cdot)$, respectively. Similarly, the augmented longitudinal views, \mathbf{x}_{g1} and \mathbf{x}_{g2} are fed into $g_q(\cdot)$ and $g_m(\cdot)$, respectively. After feature extraction and projection, the corresponding vectors, \mathbf{y}_{f1} , \mathbf{y}_{f2} , \mathbf{y}_{g1} , and \mathbf{y}_{g2} are obtained, where \mathbf{y}_{f1} and \mathbf{y}_{f2} , \mathbf{y}_{g1} and \mathbf{y}_{g2} are considered as two positive pairs since they are from the same image (\mathbf{x}_f and \mathbf{x}_g). The loss function is described in the following subsection. After computing the loss, the weights of query encoders ($f_q(\cdot)$ and $g_q(\cdot)$) are updated through back-propagation, and the weights of momentum encoders ($f_m(\cdot)$ and $g_m(\cdot)$) are updated by exponential moving average (EMA):

$$\theta_m^t \leftarrow \alpha \cdot \theta_m^{t-1} + (1 - \alpha) \cdot \theta_m^t, \quad (1)$$

where θ_q and θ_m denote the weights of the query and momentum encoders, respectively. The superscript t denotes the training step, and $\alpha \in [0, 1]$ is the momentum coefficient that controls the speed of the weight update. After weight updating, \mathbf{y}_{f2} and \mathbf{y}_{g2} are sent to the memory bank as new vectors, and the oldest vectors in the memory bank are dequeued. In our framework, the query encoders, $f_q(\cdot)$ and $g_q(\cdot)$ share the same weights, whereas the momentum encoders, $f_m(\cdot)$ and $g_m(\cdot)$ share the same weights. We also tried separate weights for the transverse and longitudinal encoders but found it better to use the weight-sharing mechanism.

3.2. Adaptive loss

Our loss function comprises two parts: single-view and cross-view contrastive losses. The single-view contrastive loss reduces the feature distance between different augmented versions of the same image, enabling the encoder

to learn transformation-invariant features. It can be written as:

$$\begin{aligned} L_{ff} &= C(y_{f1}, y_{f2}), \\ L_{gg} &= C(y_{g1}, y_{g2}), \end{aligned} \quad (2)$$

where $C(\cdot)$ is the InfoNCE loss (Oord, Li and Vinyals, 2018), which can be expressed as:

$$C(q, k) = -\log \frac{\exp(\text{sim}(q, k)/\tau)}{\exp(\text{sim}(q, k)/\tau) + \sum_{i=1}^N \exp(\text{sim}(q, t_i)/\tau)}, \quad (3)$$

where τ is a temperature parameter and $\text{sim}(\cdot)$ is the operator for similarity measurement; here, it is set as cosine similarity. The vector q and vector k are a positive pair, whereas vector q and vector t_i ($i \in \{1, 2, \dots, N\}$) are a negative pair. In this study, the vector (y_{f1} and y_{g1}) generated by query encoders ($f_q(\cdot)$ and $g_q(\cdot)$) and each vector stored in the memory bank are considered negative pairs because they are from different images or from the same image but at different training steps. Thus, there are K negative pairs and a positive pair, which forms the log loss of a $(K+1)$ -way softmax-based classifier.

The cross-view contrastive loss reduces the feature distance between the transverse and longitudinal views of the same nodule, which enables the encoder to learn view-invariant features. It can be expressed as:

$$\begin{aligned} L_{fg} &= C(y_{f1}, y_{g2}), \\ L_{gf} &= C(y_{g1}, y_{f2}). \end{aligned} \quad (4)$$

A naive idea is to combine these losses directly in the form of a summation as follows:

$$L_{pair} = L_{ff} + L_{gg} + L_{fg} + L_{gf}. \quad (5)$$

Obviously, this loss function is only used for paired views, and those unpaired views cannot be utilized. Unfortunately, it is possible to happen that one of the two views is missing in clinical practice. To solve this problem, Li et al. (2020); Taleb et al. (2017) used CycleGAN to synthesize the missing views before pre-training. However, image synthesis itself requires paired data, and synthetic images may be harmful to pre-training owing to poor synthesis quality. Therefore, we propose the following adaptive loss function:

$$L_{adaptive} = a \cdot L_{ff} + b \cdot L_{gg} + ab \cdot \lambda(L_{fg} + L_{gf}), \quad (6)$$

where $a, b \in \{0, 1\}$ are two indicator functions that evaluate zero if the corresponding view is missing, and λ is a coefficient that balances the single-view and cross-view contrastive losses. If one of the two views is missing, $L_{adaptive}$ is equal to L_{ff} or L_{gg} , which means that it adaptively becomes a single-view contrastive loss. This eliminates the limitation of requiring paired views. The final loss function is the average of the loss functions corresponding to each patient in the batch X . In summary, this adaptive loss function has two advantages: (1) The encoder can learn both transformation-invariant and view-invariant features, which is superior to only learning transformation-invariant features. (2) The encoder can be optimized on both single-view data and multi-view data, which is flexible and does not require paired data.

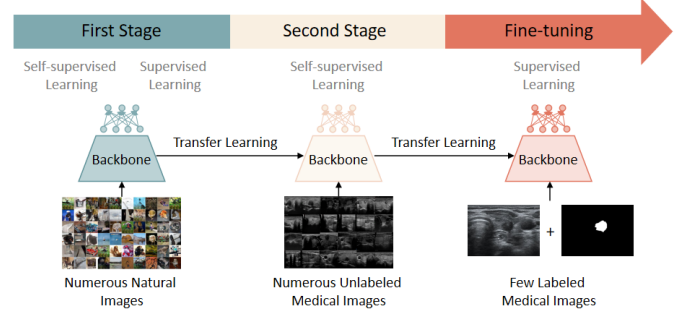


Figure 4: Two-stage pre-training. In the first stage, we train the model on ImageNet in a supervised and self-supervised learning manner. In the second stage, we first initialize the model with the learned weights from the first stage and train the model on unlabeled target medical images in a self-supervised manner. Finally, the model is fine-tuned for the target tasks.

3.3. Two-stage pre-training

For medical image analysis, self-supervised pre-training on medical images can alleviate the domain shift between ImageNet and medical datasets compared with ImageNet pre-training. However, it is also limited by the challenge of collecting numerous medical images, owing to patient privacy. In this study, instead of using one-stage pre-training, we divided the pre-training process into two stages, as shown in Fig.4. In the first stage, we trained the model on ImageNet in a supervised and self-supervised manner. This enables the model to learn the features of natural images and provides good initialization for medical image analysis. However, this ability is susceptible to domain shift and may not be effective for certain medical images (Tajbakhsh, Shin, Gurudu, Hurst, Kendall, Gotway and Liang, 2016). In the second stage, we further trained the model on unlabeled medical images similar to the target images in a self-supervised manner. This enables the model to learn the features of medical images and enhances its ability to handle such medical image analysis tasks. The pre-trained weights of common network architectures on ImageNet are usually available in the deep learning community. We can skip the first stage and directly initialize our model using these weights in the second stage. After the two-stage pre-training, we fine-tuned our model for target tasks using a small number of labeled target images in a supervised learning manner.

3.4. Implementation details

We divided the implementation into four aspects: network architecture, data augmentation, hyperparameters, and optimization settings. For the network architecture, we used ResNet50 (He, Zhang, Ren and Sun, 2016) as the backbone and a two-layer fully connected layer as the projection head. The hidden and output dimensions of the projection head are 512 and 128, respectively. For data augmentation, we employed common techniques, including cropping, resizing, brightness adjustment, contrast adjustment, Gaussian blur, horizontal flip, and rotation. For the hyperparameters, the temperature parameter τ was set to 0.1, and the memory

Table 1

Summary of the compared methods. These are the current SOTA self-supervised methods in medical image analysis. The categories include generative, contrastive, and combinations of the two. The last column refers to the images used in the original papers.

Methods	Publication	Years	Category	Images for pre-training
MoCo v2 (He et al., 2020)	CVPR	2020	contrastive	ImageNet images
C2L (Zhou et al., 2020)	MICCAI	2020	contrastive	Chest X-ray images 2D Chest CT slices
MG (Zhou et al., 2021c)	MedIA	2020	generative	3D Chest CT volumes Chest X-ray images
PCRL (Zhou, Lu, Yang, Han and Yu, 2021a)	ICCV	2021	contrastive & generative	Chest X-ray images 3D Chest CT volumes
CAiD (Taher, Haghghi, Gotway and Liang, 2022)	MIDL	2022	contrastive & generative	Chest X-ray images
DiRA (Haghghi, Taher, Gotway and Liang, 2022)	CVPR	2022	contrastive & generative	Chest X-ray images 3D Chest CT volumes
SSFL (Li et al., 2020)	TMI	2020	contrastive	Multimodal fundus images

bank size K was set to 512. The momentum coefficient α was set to 0.99 and the loss coefficient λ was set to 0.5. For the optimization settings, we used the following details: SGD optimizer, cosine learning rate decay scheduler, initial learning rate of 0.03, weight decay of 0.0001, batch size of 128 (64 patients per iteration), and 200 training epochs. The intensity normalization (i.e. subtract the mean and divide by the standard deviation) was performed. To demonstrate that the target tasks benefit from our method rather than from special tricks, we also implemented MoCo v2 using the above settings. For the two-stage pre-training, we modified the memory bank size K to 1024 and the initial learning rate to 0.01. We obtained the supervised ImageNet pre-trained weights from PyTorch official repository and self-supervised ImageNet pre-trained weights from MMSelfSup Contributors (2021). All experiments in this study were performed on a PyTorch platform using servers equipped with NVIDIA RTX A40 GPUs.

4. Materials and Experiments

In this section, we introduce our large-scale thyroid ultrasound dataset collected from multiple centers, compared methods, and target tasks.

4.1. Dataset

To evaluate the proposed method, we constructed a large-scale thyroid ultrasound dataset. Our dataset has the following characteristics: (1) It was collected from multiple centers consisting of more than 20 hospitals and sites. These centers are located in different regions of China, such as Shanghai, Chengdu in Sichuan, and Changzhou in Jiangsu, providing regional diversity and physician scanning diversity. (2) Images were generated using more than 30 types of ultrasound imaging equipment under different settings. The equipment primarily included the Esaote Mylab series, GE LOGIQ E9, Mindray Resona7 series, Philips EPIQ7, SIEMENS ACUSONS 2000, SAMSUNG RS80A, and TOSHIBA Aplio series, which provides diverse imaging equipment. (3) This dataset contains 5224 patients ranging in

age from 9 to 82 years old, providing patient diversity. This diverse dataset ensured the reliability of our experimental results.

Our dataset contains 2216 patients with benign nodules and 3008 patients with malignant nodules. All nodule labels were determined using FNA biopsy reports. For the ultrasound images with multiple nodules, the labels were annotated as malignant if one nodule was malignant. This is the largest multicenter thyroid ultrasound image dataset with pathological labels. Additionally, the nodule masks were annotated by experienced physicians. Only the largest nodule in the image was labeled. For preprocessing, we first used the largest connected component algorithm to obtain the largest connected region. Second, we cut the surrounding regions to extract it as the region of interest (ROI). Third, the extracted ROIs were cropped or padded and then resized to 256×256 pixels. Finally, we manually verified all images and corrected the errors.

Our dataset contains 9669 images from 5224 patients. It is divided into two groups. The first group includes 779 patients with a single transverse or longitudinal view. The second group contains 4445 patients with both transverse and longitudinal views. Patients in the second group were randomly and evenly divided into 10 subsets. We first established a fixed test set by randomly selecting two subsets. The remaining 8 subsets are used for cross-validation, with a training set to validation set ratio of 7:1. We selected 5 sets of training-validation sets for cross-validation. The images in second group were used for target task fine-tuning, and all metrics in this study were reported based on the fixed test set. The training and validation set of the second group and the first group were used for pre-training. As a result, the model never sees the test set in the pre-training stage, which avoids feature memorization or any form of information leakage. To comprehensively evaluate the proposed method, we further divided the training set into different proportions: 10%, 20%, and 50%, containing 310, 621, and 1554 patients, respectively. We investigated the target performance with limited training images using different proportions of data.

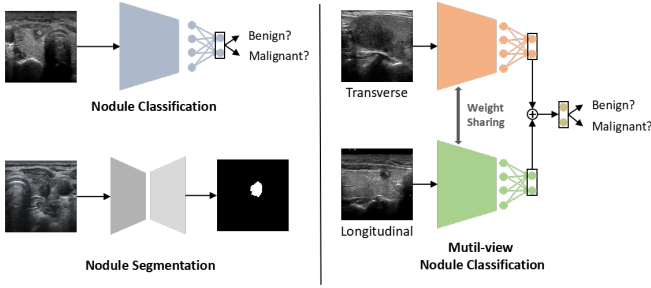


Figure 5: Networks of three target tasks. For NC, we use ResNet50 as the backbone and a one-layer fully connected layer as the classifier. For NS, we use the UNet as the network. For MNC, both two views have a network that consists of ResNet50 and a one-layer fully connected layer, and the two networks share the same weights.

4.2. Compared methods

We compared the proposed method with recent self-supervised methods on other images, including MoCo v2 (He et al., 2020), C2L (Zhou et al., 2020), MG (Zhou et al., 2021c), PCRL (Zhou et al., 2021a), CAiD (Taher et al., 2022), DiRA (Haghighi et al., 2022), and SSFL (Li et al., 2020). As shown in Table.1, these methods were published in reputed journals or conferences and are current state-of-the-art (SOTA) self-supervised methods. These methods were re-implemented on our ultrasound images by referring to the authors' codes and articles. For a fair comparison, the original methods were uniformly modified while maintaining their specificity.

4.3. Target tasks

We employed three target tasks to evaluate the effectiveness of pre-training: nodule classification (NC), nodule segmentation (NS), and multi-view nodule classification (MNC). These networks are illustrated in Fig.5. For NC, we trained a model to predict benign and malignant nodules. We did not distinguish between the transverse and longitudinal images and treated them as independent samples. We used cross-entropy loss as the loss function. For NS, we adopted UNet (Ronneberger, Fischer and Brox, 2015) as the segmentation network. The encoder was ResNet50 and the decoder consisted of multiple convolutional layers and interpolation operations. We used Dice loss as the loss function. For MNC, we treated the transverse and longitudinal images of the same patient as the joint sample. The network consisted of two branches, corresponding to the two views. Each branch consists of ResNet50 and MLP. Both branches shared the same weights. This design suppresses overfitting by reducing the number of parameters, and its effectiveness has been proven (Wang et al., 2020; Huang, Dong, Jia, Zhou, Ni, Cheng and Huang, 2022). We used the average of the outputs of the two branches as the final prediction results. We calculated the cross-entropy loss for the transverse branch, longitudinal branch, and final prediction, and we adopted the sum of the three losses as the final loss function.

Table 2

NC results (Unit:%). The random model was trained from scratch with an initial learning rate of 0.02 and 200 training epochs. Other models were trained with an initial learning rate of 0.005.

Init	$r=10\%$	$r=20\%$	$r=50\%$	$r=100\%$
Random	60.28	69.16	78.60	84.38
MoCo v2	<u>76.30</u>	<u>78.40</u>	82.78	85.52
C2L	75.86	78.21	82.84	86.01
MG	54.86	66.89	74.70	81.02
PCRL	74.02	77.18	83.42	85.10
CAiD	74.30	74.74	82.58	85.21
DiRA	75.87	78.19	<u>83.59</u>	<u>86.28</u>
SSFL	75.95	78.17	<u>83.02</u>	85.95
Ours	77.75*	79.87*	84.36*	86.30

Table 3

NS results (Unit:%). The random model was trained from scratch with an initial learning rate of 0.02 and 100 training epochs. Other models were trained with an initial learning rate of 0.01.

Init	$r=10\%$	$r=20\%$	$r=50\%$	$r=100\%$
Random	68.11	79.60	84.56	86.36
MoCo v2	76.27	80.58	<u>84.72</u>	<u>86.46</u>
C2L	75.28	80.42	84.17	86.01
MG	66.99	78.38	83.97	86.17
PCRL	75.63	80.45	84.57	86.46
CAiD	73.08	78.58	83.39	85.55
DiRA	75.06	79.98	83.90	85.99
SSFL	<u>77.71</u>	<u>81.76</u>	84.66	86.44
Ours	79.69*	82.31*	84.89	86.50

We used the same data augmentation for the three target tasks: horizontal flip, brightness and contrast adjustment, random scale, rotation, and translation. The SGD optimizer and cosine learning rate decay scheduler were employed to optimize all models. We used an early stopping mechanism on the validation set to avoid overfitting. We set the weight decay to 0.00001 and the batch size to 32. For the randomly initialized models, we used the initial learning rate $\in \{0.01, 0.02, 0.05\}$ and training epochs $\in \{100, 200\}$ to search for the best results as the baseline. For models using pre-trained weights, we used the initial learning rate $\in \{0.005, 0.01\}$ and 100 epochs to train the models. We used the area under curve (AUC) score to evaluate nodule classification and the Dice score to evaluate nodule segmentation. Both pre-training and fine-tuning were performed five times.

5. Results

In this section, we first compare the proposed method with SOTA methods and then compare the two-stage pre-training with ImageNet pre-training. For convenience, randomly initialized models are denoted as "Random". "SPIN" and "SSPIN" denote supervised and self-supervised pre-training on ImageNet, respectively. The two-stage pre-training

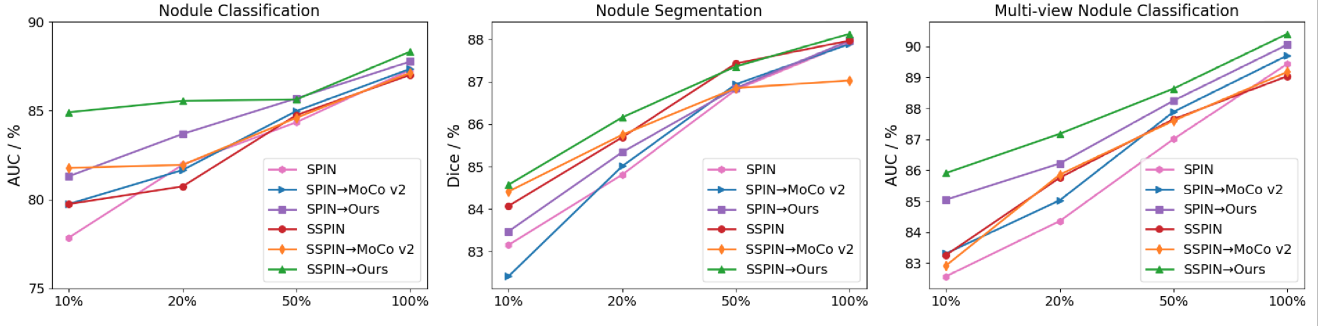


Figure 6: Comparison between ImageNet pre-training and two-stage pre-training, based on three target tasks and different proportions of training data.

Table 4

MNC results (Unit:%). The random model was trained from scratch with an initial learning rate of 0.01 and 200 training epochs. Other models were trained with an initial learning rate of 0.005.

Init	$r=10\%$	$r=20\%$	$r=50\%$	$r=100\%$
Random	58.90	66.80	78.37	86.69
MoCo v2	77.28	80.36	85.94	88.23
C2L	<u>77.43</u>	80.33	85.65	88.87
MG	55.75	62.43	72.32	81.69
PCRL	73.42	78.87	85.60	88.30
CAiD	74.89	78.76	83.46	87.69
DiRA	76.39	78.27	<u>86.70</u>	<u>89.33</u>
SSFL	74.41	<u>81.06</u>	86.08	88.27
Ours	78.17*	82.45*	87.08	89.48

is expressed in the form of “A→B”. We use r to represent the proportion of accessible data in the entire training set. Quantitative results are reported as the mean of five trials. The best and second-best results are bolded and underlined, respectively. Paired samples t -test was performed.

5.1. Nodule classification

Table.2 shows the results. The p -values between our results and the second-best results were calculated. The symbol ‘*’ indicates p -value <0.001 and is considered significant. The random model shows a large variance in performance when different proportions of the training data are used for training. Specifically, the random model achieved average AUC scores of 60.28%, 69.16%, 78.60%, and 84.38% with different proportions of training data (10%, 20%, 50%, and 100%), respectively. From 10% to 100%, the AUC score increased by 24.1%. This shows that the random model is highly sensitive to the amount of training data, and insufficient data causes the model to perform poorly. In addition to the MG model, other self-supervised models have improved classification performance over the random model, and our model achieved the largest boost. Compared with the random model, our model improved the AUC scores by 17.47%, 10.71%, 5.76%, and 1.92%, respectively. Our model also significantly outperformed all

other self-supervised models with limited manual labels, demonstrating its effectiveness.

5.2. Nodule segmentation

Table.3 presents the nodule segmentation results. The performance of the random model varies significantly for different proportions of training data. This indicates that the random model performs poorly on a few training datasets. In contrast, using less training data, self-supervised models other than the MG model improve segmentation performance. Our model comprehensively outperforms other self-supervised models and achieves average Dice scores of 79.69%, 82.31%, 84.89%, and 86.50%. This shows that our method can also significantly improve nodule segmentation with limited manual labels, demonstrating its effectiveness.

5.3. Multi-view nodule classification

Table.4 presents the multi-view nodule classification results. Similarly, for different proportions of training data, the random model exhibits a large variance in performance. The score difference between 10% and 100% of the training data is 27.79%. Compared with the random model, the self-supervised models improve the multi-view nodule classification performance, except for the MG model. The DiRA model, for example, improves the AUC scores by 17.49%, 11.47%, 8.33%, and 2.64%. This is a major improvement, particularly with less training data. Our model achieves average AUC scores of 78.17%, 82.45%, 87.08%, and 89.48%, which significantly outperforms other self-supervised models when using a small amount of training data (i.e. $r=10\%$ and 20%). This demonstrates the effectiveness of the proposed method.

5.4. Two-stage pre-training

The nodule classification, segmentation, and multi-view nodule classification results are presented in Tables.5, 6, and 7, respectively. The p -values between two-stage pre-training of our method and ImageNet pre-training were calculated. The symbol ‘†’ indicates p -value <0.001 and is considered significant. To present the results better, we drew line charts for the three tasks, as shown in Fig.6. Compared to ImageNet pre-training, the two-stage pre-training of our method almost

Table 5

NC (two-stage) results (Unit:%). All models were trained with an initial learning rate of 0.005.

Init	$r=10\%$	$r=20\%$	$r=50\%$	$r=100\%$
SPIN	77.86	81.96	84.35	87.29
SPIN→MoCo v2	79.75	81.66	84.98	87.35
SPIN→Ours	81.31[†]	83.69[†]	85.69[†]	87.76
SSPIN	79.75	80.74	84.75	87.01
SSPIN→MoCo v2	81.78	81.95	84.60	87.13
SSPIN→Ours	84.91[†]	85.55[†]	85.63[†]	88.32[†]

Table 6

NS (two-stage) results (Unit:%). All models were trained with an initial learning rate of 0.01.

Init	$r=10\%$	$r=20\%$	$r=50\%$	$r=100\%$
SPIN	83.15	84.81	86.81	87.92
SPIN→MoCo v2	82.41	85.01	86.94	87.89
SPIN→Ours	83.47	85.34[†]	86.85	87.97
SSPIN	84.07	85.69	87.43	87.97
SSPIN→MoCo v2	84.41	85.75	86.85	87.03
SSPIN→Ours	84.57[†]	86.16[†]	87.36	88.13

Table 7

MNC (two-stage) results (Unit:%). All models were trained with an initial learning rate of 0.005.

Init	$r=10\%$	$r=20\%$	$r=50\%$	$r=100\%$
SPIN	82.57	84.36	87.01	89.44
SPIN→MoCo v2	83.31	85.03	87.89	89.71
SPIN→Ours	85.05[†]	86.22[†]	88.25[†]	90.06[†]
SSPIN	83.26	85.76	87.65	89.04
SSPIN→MoCo v2	82.92	85.86	87.60	89.18
SSPIN→Ours	85.91[†]	87.18[†]	88.64[†]	90.41[†]

always improves performance. For example, the AUC scores of SPIN→Ours are 3.45%, 1.73%, 1.34%, and 0.47% higher than SPIN in NC. In NS, SSPIN→Ours is only slightly lower than SSPIN when $r=50\%$, and outperforms SSPIN in other proportions. In MNC, SPIN→Ours improved the AUC scores by 2.48%, 1.86%, 1.24%, and 0.62%, respectively. Compared to ImageNet pre-training, MoCo v2's two-stage pre-training improves performance or reaches competitive performance. Two-stage pre-training of our method outperforms that of MoCo v2 in most cases, demonstrating the effectiveness of the proposed method.

5.5. Ablation study

Table.8 lists the results of three downstream tasks with different lambda values in our method. When lambda is equal to 0, the method degenerates to MoCo v2. This model performs worst in given lambda values. This demonstrates the effectiveness of cross-view contrastive learning. Our method achieves the best performance when lambda is equal to 0.5. We also verify the impact of the number of paired

Table 8

Results (Unit:%) of three downstream tasks with different lambda values in our method.

Tasks	r	λ in Equation 6			
		0	0.2	0.5	1.0
NC	10%	76.30	77.67	77.75	76.79
	20%	78.40	79.71	79.87	79.23
NS	10%	76.27	78.89	79.69	79.56
	20%	80.58	81.97	82.31	82.43
MNC	10%	77.28	77.52	78.17	77.47
	20%	80.36	81.79	82.45	82.29

Table 9

NC results (Unit:%). The dataset used for pre-training consists of all paired images (training and validation set in the second group) and different proportions of 779 unpaired views.

r	Init	+ (%) of 779 unpaired views			
		0	20%	50%	100%
10%	MoCo v2	75.02	75.14	75.67	76.30
	Ours	77.13	77.20	77.54	77.75
20%	MoCo v2	76.08	76.46	78.41	78.40
	Ours	79.33	79.23	79.79	79.87
50%	MoCo v2	81.34	81.83	81.94	82.78
	Ours	83.76	84.04	84.29	84.36
10%	MoCo v2	85.37	85.33	85.45	85.52
	Ours	86.14	86.34	86.30	86.50

views on cross-view contrastive loss. The results are shown in Table.9. From the table we can observe: (1) Our method always outperforms MoCo v2 even though MoCo v2 uses all unpaired views while our method does not use any unpaired views. This shows that cross-view contrastive loss is important. (2) Our approach continues to improve performance as available unpaired views increase. This demonstrates the effectiveness of the proposed adaptive loss function. Considering the clinical practice of coexistence of multiple views and missing views in thyroid ultrasound examination, our method can help improve thyroid ultrasound diagnosis with limited labeled data.

6. Discussions

6.1. Why is our method better?

Our method significantly outperformed those designed for single-view images. This could be because our pre-training method makes the model pay more attention to the nodule area, which provides a good prior for the three target tasks. To verify this, we used activation maps as nodules' segmentation maps and computed the Dice score using nodule masks. Specifically, we froze the pre-trained ResNet50 and fed all images from the test set into it. We obtained the feature maps before the global pooling layer. The activation map is obtained by directly averaging the feature maps along the channel dimension. We resized the activation map to the

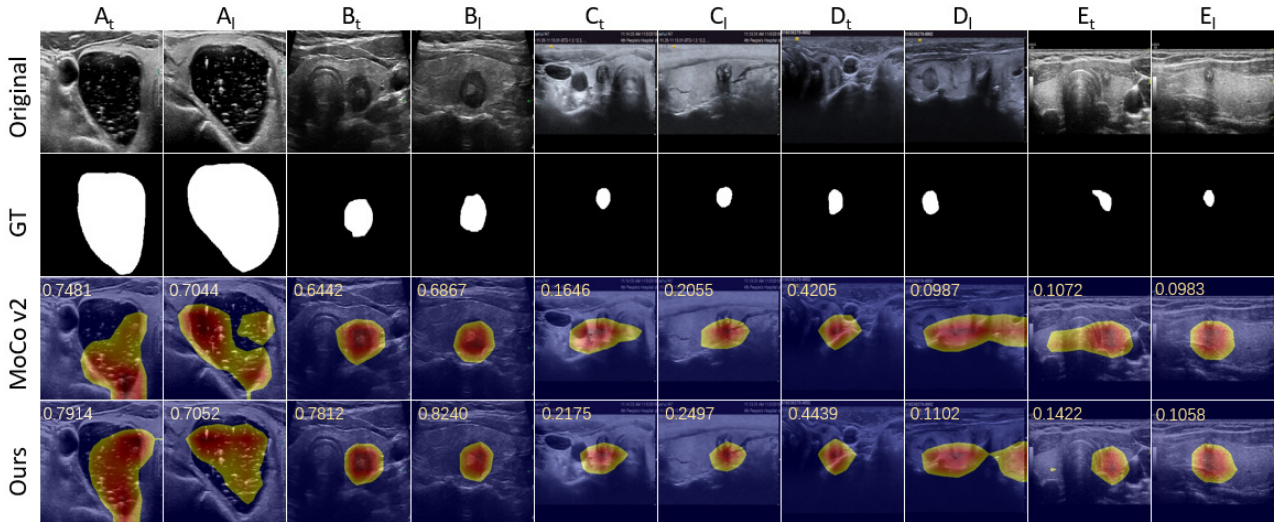


Figure 7: Activation map visualization. The first row is the original image, and the second row is the corresponding mask. The third and fourth rows are the activation maps of MoCo v2 and our method, respectively. Numbers on the activation map represent Dice scores. We show five pairs of images and use a threshold of 0.5.

Table 10

Average Dice score between the activation maps and nodule masks at different thresholds (Unit:%).

Pre-training	$t=0.3$	$t=0.4$	$t=0.5$	$t=0.6$	$t=0.7$
MoCo v2	32.01	34.61	35.93	35.86	33.75
Ours	36.42	39.50	41.11	40.96	38.20

original image size and normalized it to $[0,1]$. We obtained the segmentation map by binarizing the activation map with different thresholds $t \in \{0.3, 0.4, 0.5, 0.6, 0.7\}$. We compared our method with MoCo v2. The quantitative results are presented in Table.10 and the qualitative visualizations are shown in Fig.7. Our method always has higher Dice scores than MoCo v2 and pays more attention to the nodule area, although nodule sizes vary significantly. Several studies have indicated that lesion segmentation facilitates accurate disease classification (Zhang, Tang, Cao, Han, Xiao, Ma and Chang, 2021; Zhou, Chen, Li, Liu, Xu, Wang, Yap and Shen, 2021b). This may explain why our method achieves better classification and segmentation performance. In addition, our method also outperforms SSFL (Li et al., 2020), which also adopts multi-view contrastive learning. This is because SSFL can only utilize paired data, and images with only one transverse or one longitudinal view are not utilized. Overall, our method benefits from multi-view contrastive learning that eliminates the paired data constraints.

6.2. Why is two-stage pre-training better?

The two-stage pre-training uses ImageNet pre-training and significantly surpasses it without using additional labels. Good pre-trained weights provide more reusable features (Neyshabur, Sedghi and Zhang, 2020). Following (Neyshabur et al., 2020), we evaluated the degree of feature reuse by measuring the feature similarity of the different

Table 11

Comparison of feature reuse between two-stage pre-training and ImageNet pre-training. Each row presents the CKA score for different intermediate layers before and after fine-tuning models in nodule classification.

Pre-training	conv1	layer1	layer2	layer3	layer4
SPIN	0.966	0.920	0.965	0.911	0.114
SPIN→Ours	0.993	0.993	0.991	0.846	0.179
SSPIN	0.942	0.947	0.957	0.854	0.126
SSPIN→Ours	0.975	0.989	0.982	0.937	0.381

layers of the models before and after fine-tuning using centered kernel alignment (CKA) (Kornblith, Norouzi, Lee and Hinton, 2019). Fig.8 shows the visualization results. A higher CKA score indicates more feature reuse, and we primarily focused on the CKA scores at the diagonal positions. Two-stage models have higher CKA scores than corresponding ImageNet pre-training models, especially in the low/mid-layer. This shows that the two-stage pre-training provides more reusable features than the ImageNet pre-training. We further computed the CKA scores of several representative layers, as shown in Table.11. The two-stage model corresponding to SSPIN improves the CKA scores across the board more significantly in the highest layer. It is well known that in CNNs, the lower layers extract detailed features, while the higher layers extract task-specific features. This shows that the two-stage self-supervised pre-training not only provides more reusable detailed features but also provides more reusable task-specific features. This is probably why it performs better than ImageNet pre-training, even without using any additional labels.

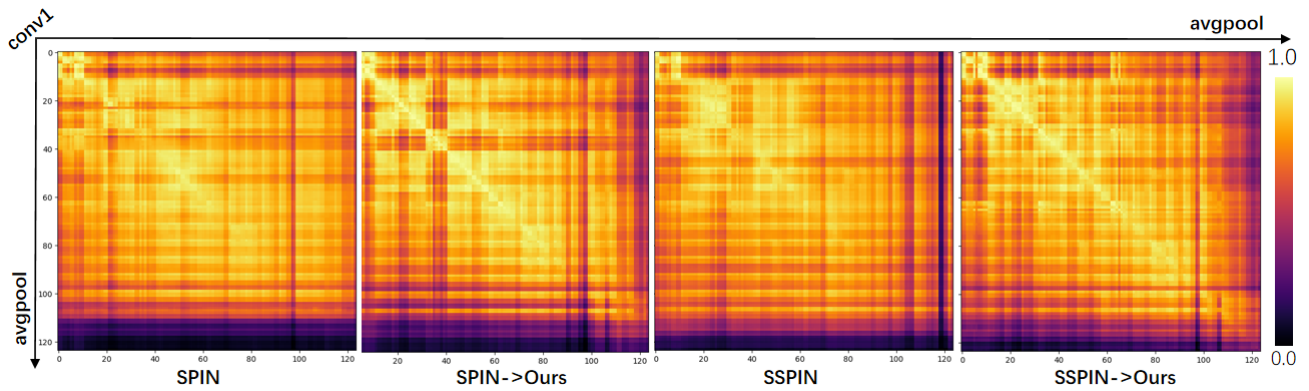


Figure 8: Centered kernel alignment (CKA) score map. Based on nodule classification, we calculate the CKA scores of the different layers of models before and after fine-tuning, including convolutional, batchnorm, and pooling layers. The value at coordinates (i, j) in each map represents the CKA score between the i -th layer of the model before fine-tuning the j -th layer of the model after fine-tuning.

7. Conclusion

We proposed a multi-view contrastive self-supervised method to improve the nodule classification and segmentation performance of thyroid ultrasound images with limited manual labels. Our method enables the model to learn transformation- and view-invariant features. To address the issue of missing views, we designed an adaptive loss function that eliminates the need for paired views. We also adopted a two-stage pre-training strategy to alleviate the domain shift between natural and medical images. To verify the effectiveness of the proposed method, we constructed a large-scale thyroid ultrasound image dataset from more than 20 hospitals. The results of the extensive experiments show that the proposed method significantly improves nodule classification and segmentation performance compared to random initialization and outperforms other SOTA self-supervised methods with limited manual labels. The results also show that the two-stage pre-training strategy can significantly boost the target performance.

Conflict of interest statement

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Natural Science Innovative Research Group Project (61821002), the Key Project of the National Natural Science Foundation of China (51832001), and the Frontier Fundamental Research Program of Jiangsu Province for Leading Technology (BK20222002).

References

Azizi, S., Mustafa, B., Ryan, F., Beaver, Z., Freyberg, J., Deaton, J., Loh, A., Karthikesalingam, A., Kornblith, S., Chen, T., et al., 2021.

Big self-supervised models advance medical image classification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3478–3488.

Basu, S., Singla, S., Gupta, M., Rana, P., Gupta, P., Arora, C., 2022. Unsupervised contrastive learning of image representations from ultrasound videos with hard negative mining, in: Medical Image Computing and Computer Assisted Intervention—MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part IV, Springer. pp. 423–433.

Bethesda, M., 2018. Seer cancer stat facts thyroid cancer. National Cancer Institute. [accessed on 10 May 2021].

Chen, J., You, H., Li, K., 2020a. A review of thyroid gland segmentation and thyroid nodule segmentation methods for medical ultrasound images. *Computer methods and programs in biomedicine* 185, 105329.

Chen, L., Bentley, P., Mori, K., Misawa, K., Fujiwara, M., Rueckert, D., 2019. Self-supervised learning for medical image analysis using image context restoration. *Medical image analysis* 58, 101539.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020b. A simple framework for contrastive learning of visual representations, in: International conference on machine learning, PMLR. pp. 1597–1607.

Chen, X., Fan, H., Girshick, R., He, K., 2020c. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*.

Contributors, M., 2021. MMSelfSup: Openmmlab self-supervised learning toolbox and benchmark. <https://github.com/open-mmlab/mmselfsup>.

Deng, P., Han, X., Wei, X., Chang, L., 2022. Automatic classification of thyroid nodules in ultrasound images using a multi-task attention network guided by clinical knowledge. *Computers in Biology and Medicine* 150, 106172.

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

Fedorov, A., Sylvain, T., Geenjaer, E., Luck, M., Wu, L., DeRamus, T.P., Kirilin, A., Bleklov, D., Calhoun, V.D., Plis, S.M., 2021a. Self-supervised multimodal domino: in search of biomarkers for alzheimer's disease, in: 2021 IEEE 9th International Conference on Healthcare Informatics (ICHI), IEEE. pp. 23–30.

Fedorov, A., Wu, L., Sylvain, T., Luck, M., DeRamus, T.P., Bleklov, D., Plis, S.M., Calhoun, V.D., 2021b. On self-supervised multimodal representation learning: an application to alzheimer's disease, in: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), IEEE. pp. 1548–1552.

Gidaris, S., Singh, P., Komodakis, N., 2018. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*.

Gong, H., Chen, J., Chen, G., Li, H., Li, G., Chen, F., 2023. Thyroid region prior guided attention for ultrasound segmentation of thyroid nodules.

- Computers in Biology and Medicine 155, 106389.
- Haghighi, F., Taher, M.R.H., Gotway, M.B., Liang, J., 2022. Dira: discriminative, restorative, and adversarial learning for self-supervised medical image analysis, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 20824–20834.
- Hassani, K., Khasahmadi, A.H., 2020. Contrastive multi-view representation learning on graphs, in: International conference on machine learning, PMLR. pp. 4116–4126.
- He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R., 2022. Masked autoencoders are scalable vision learners, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16000–16009.
- He, K., Fan, H., Wu, Y., Xie, S., Girshick, R., 2020. Momentum contrast for unsupervised visual representation learning, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 9729–9738.
- He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.
- Hervella, Á.S., Rouco, J., Novo, J., Ortega, M., 2021. Self-supervised multimodal reconstruction pre-training for retinal computer-aided diagnosis. *Expert Systems with Applications* 185, 115598.
- Hervella, I.S., Rouco, J., Novo, J., Ortega, M., 2020. Self-supervised multimodal reconstruction of retinal images over paired datasets. *Expert Systems with Applications* 161, 113674.
- Huang, H., Dong, Y., Jia, X., Zhou, J., Ni, D., Cheng, J., Huang, R., 2022. Personalized diagnostic tool for thyroid cancer classification using multi-view ultrasound, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part III, Springer. pp. 665–674.
- Kang, Q., Lao, Q., Li, Y., Jiang, Z., Qiu, Y., Zhang, S., Li, K., 2022. Thyroid nodule segmentation and classification in ultrasound images through intra-and inter-task consistent learning. *Medical image analysis* 79, 102443.
- Kim, D., Song, B.C., 2022. Emotion-aware multi-view contrastive learning for facial emotion recognition, in: European Conference on Computer Vision, Springer. pp. 178–195.
- Kornblith, S., Norouzi, M., Lee, H., Hinton, G., 2019. Similarity of neural network representations revisited, in: International Conference on Machine Learning, PMLR. pp. 3519–3529.
- Li, X., Jia, M., Islam, M.T., Yu, L., Xing, L., 2020. Self-supervised feature learning via exploiting multi-modal data for retinal disease diagnosis. *IEEE Transactions on Medical Imaging* 39, 4023–4033.
- Li, Y., Yang, M., Zhang, Z., 2018. A survey of multi-view representation learning. *IEEE transactions on knowledge and data engineering* 31, 1863–1883.
- Liu, J., Dong, B., Wang, S., Cui, H., Fan, D.P., Ma, J., Chen, G., 2021. Covid-19 lung infection segmentation with a novel two-stage cross-domain transfer learning framework. *Medical image analysis* 74, 102205.
- Meng, J., Tan, Z., Yu, Y., Wang, P., Liu, S., 2022. Tl-med: A two-stage transfer learning recognition model for medical images of covid-19. *Biocybernetics and Biomedical Engineering* 42, 842–855.
- Neyshabur, B., Sedghi, H., Zhang, C., 2020. What is being transferred in transfer learning? *Advances in neural information processing systems* 33, 512–523.
- Noroozi, M., Favaro, P., 2016. Unsupervised learning of visual representations by solving jigsaw puzzles, in: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VI, Springer. pp. 69–84.
- Oord, A.v.d., Li, Y., Vinyals, O., 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.
- Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A., 2016. Context encoders: Feature learning by inpainting, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2536–2544.
- Punn, N.S., Agarwal, S., 2022. Bt-UNET: A self-supervised learning framework for biomedical image segmentation using barlow twins with u-net models. *Machine Learning*, 1–16.
- Raghu, M., Zhang, C., Kleinberg, J., Bengio, S., 2019. Transfusion: Understanding transfer learning for medical imaging. *Advances in neural information processing systems* 32.
- Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18, Springer. pp. 234–241.
- Roy, S., Etemad, A., 2021. Self-supervised contrastive learning of multi-view facial expressions, in: Proceedings of the 2021 International Conference on Multimodal Interaction, pp. 253–257.
- Shah, K., Shah, A., Lau, C.P., de Melo, C.M., Chellappa, R., 2023. Multi-view action recognition using contrastive learning, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 3381–3391.
- Sharifi, Y., Bakhshali, M.A., Dehghani, T., DanaiAshgzi, M., Sargolzaei, M., Eslami, S., 2021. Deep learning on ultrasound images of thyroid nodules. *Biocybernetics and Biomedical Engineering* 41, 636–655.
- Shurrah, S., Duwairi, R., 2022. Self-supervised learning methods and applications in medical imaging analysis: A survey. *PeerJ Computer Science* 8, e1045.
- Smith-Bindman, R., Miglioretti, D.L., Johnson, E., Lee, C., Feigelson, H.S., Flynn, M., Greenlee, R.T., Kruger, R.L., Hornbrook, M.C., Roblin, D., et al., 2012. Use of diagnostic imaging studies and associated radiation exposure for patients enrolled in large integrated health care systems, 1996–2010. *Jama* 307, 2400–2409.
- Sun, J., Wu, B., Zhao, T., Gao, L., Xie, K., Lin, T., Sui, J., Li, X., Wu, X., Ni, X., 2023. Classification for thyroid nodule using vit with contrastive learning in ultrasound images. *Computers in Biology and Medicine* 152, 106444.
- Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., Bray, F., 2021. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians* 71, 209–249.
- Taher, M.R.H., Haghighi, F., Gotway, M.B., Liang, J., 2022. Caid: Context-aware instance discrimination for self-supervised learning in medical imaging, in: International Conference on Medical Imaging with Deep Learning, PMLR. pp. 535–551.
- Tajbakhsh, N., Shin, J.Y., Gurudu, S.R., Hurst, R.T., Kendall, C.B., Gotway, M.B., Liang, J., 2016. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging* 35, 1299–1312.
- Taleb, A., Lippert, C., Klein, T., Nabi, M., 2017. Self-supervised learning for medical images by solving multimodal jigsaw puzzles. *Ieee Transactions on Medical Imaging* 12729, 661–673.
- Verma, A., Tapaswi, M., 2022. Can we adopt self-supervised pretraining for chest x-rays? *arXiv preprint arXiv:2211.12931*.
- Wang, J., Miao, J., Yang, X., Li, R., Zhou, G., Huang, Y., Lin, Z., Xue, W., Jia, X., Zhou, J., et al., 2020. Auto-weighting for breast cancer classification in multimodal ultrasound, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VI 23, Springer. pp. 190–199.
- Wu, Y., Xie, R., Zhu, Y., Ao, X., Chen, X., Zhang, X., Zhuang, F., Lin, L., He, Q., 2022. Multi-view multi-behavior contrastive learning in recommendation, in: International Conference on Database Systems for Advanced Applications, Springer. pp. 166–182.
- Xiang, Z., Zhuo, Q., Zhao, C., Deng, X., Zhu, T., Wang, T., Jiang, W., Lei, B., 2022. Self-supervised multi-modal fusion network for multi-modal thyroid ultrasound image diagnosis. *Computers in Biology and Medicine* 150, 106164.
- Zhang, X., Chen, Z., Gao, J., Huang, W., Li, P., Zhang, J., 2022. A two-stage deep transfer learning model and its application for medical image processing in traditional chinese medicine. *Knowledge-Based Systems* 239, 108060.

- Zhang, Y., Tang, Y., Cao, Z., Han, M., Xiao, J., Ma, J., Chang, P., 2021. Bi-rads classification of calcification on mammograms, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VII 24, Springer. pp. 119–128.
- Zhou, H.Y., Lu, C., Yang, S., Han, X., Yu, Y., 2021a. Preservational learning improves self-supervised medical image models by reconstructing diverse contexts, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3499–3509.
- Zhou, H.Y., Yu, S., Bian, C., Hu, Y., Ma, K., Zheng, Y., 2020. Comparing to learn: Surpassing imagenet pretraining on radiographs by comparing image representations, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23, Springer. pp. 398–407.
- Zhou, Y., Chen, H., Li, Y., Liu, Q., Xu, X., Wang, S., Yap, P.T., Shen, D., 2021b. Multi-task learning for segmentation and classification of tumors in 3d automated breast ultrasound images. *Medical Image Analysis* 70, 101918.
- Zhou, Z., Sodha, V., Pang, J., Gotway, M.B., Liang, J., 2021c. Models genesis. *Medical image analysis* 67, 101840.
- Zhu, J., Li, Y., Hu, Y., Ma, K., Zhou, S.K., Zheng, Y., 2020. Rubik's cube+: A self-supervised feature learning framework for 3d medical image analysis. *Medical image analysis* 64, 101746.
- Zhu, J.Y., Park, T., Isola, P., Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE international conference on computer vision, pp. 2223–2232.