Computers in Biology and Medicine 165 (2023) 107337

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Self-supervised dual-head attentional bootstrap learning network for prostate cancer screening in transrectal ultrasound images

Xu Lu [a,b,c], Xiangjun Liu [a], Zhiwei Xiao [a], Shulian Zhang [a], Jun Huang [d], Chuan Yang [d], Shaopeng Liu [a,*]

[a] *Guangdong Polytechnic Normal University, Guangzhou 510665, China*
[b] *Guangdong Provincial Key Laboratory of Intellectual Property & Big Data, Guangzhou 510665, China*
[c] *Pazhou Lab, Guangzhou 510330, China*
[d] *Department of Ultrasonography, The First Affiliated Hospital of Jinan University, Guangzhou 510630, China*

ARTICLE INFO

Keywords:
Ultrasound imaging of the prostate
Classification
Self-supervision
Contrast learning
Parallax-attention mechanism

ABSTRACT

Current convolutional neural network-based ultrasound automatic classification models for prostate cancer often rely on extensive manual labeling. Although Self-supervised Learning (SSL) have shown promise in addressing this problem, those data that from medical scenarios contains intra-class similarity conflicts, so using loss calculations directly that include positive and negative sample pairs can mislead training. SSL method tends to focus on global consistency at the image level and does not consider the internal informative relationships of the feature map. To improve the efficiency of prostate cancer diagnosis, using SSL method to learn key diagnostic information in ultrasound images, we proposed a self-supervised dual-head attentional bootstrap learning network (SDABL), including Online-Net and Target-Net. Self-Position Attention Module (SPAM) and adaptive maximum channel attention module (CAAM) are inserted in both paths simultaneously. They captures position and inter-channel attention and of the original feature map with a small number of parameters, solve the information optimization problem of feature maps in SSL. In loss calculations, we discard the construction of negative sample pairs, and instead guide the network to learn the consistency of the location space and channel space by drawing closer to the embedding representation of positive samples continuously. We conducted numerous experiments on the prostate Transrectal ultrasound (TRUS) dataset, experiments show that our SDABL pre-training method has significant advantages over both mainstream contrast learning methods and other attention-based methods. Specifically, the SDABL pre-trained backbone achieves 80.46% accuracy on our TRUS dataset after fine-tuning.

1. Introduction

Prostate cancer is a complex disease which affects millions of men worldwide. According to 2022 cancer statistics [1], it has become the most common cancer among men in the United States and has become increasingly prevalent among Chinese men. Early diagnosis is essential for successful treatment of prostate cancer [2]. Transrectal Ultrasonography imaging is commonly used in the initial diagnosis of prostate cancer because of its low cost, rapid detection, non-invasiveness, and non-radiation [3]. In recent years, deep learning methods, especially convolutional neural networks (CNNs) [4], have achieved great success in computer-aided diagnosis (CAD). CNNs can learn high-dimensional features from medical images and have demonstrated excellent performance on several tasks such as cancer diagnosis [5] and lesion localization [6]. The success of CNNs is mainly attributed to their ability to extract representations from different dimensions of image information. However, it usually requires the support of a large-scale annotated training dataset. Clinically experienced physicians must manually label each image a time-consuming and laborious process to ensure the reliability of the dataset. Because the ultrasound images are often of low resolution, manual marking often produces errors.

One alternative to manually labeled data is to learn feature representations directly from unlabeled image data using an unsupervised scheme. Self-supervised learning (SSL) methods [7–10] exhibit similar performance to supervised learning methods on the downstream task while only using limited labeling in the initialized CNNs. Recently, Contrast Learning (CL) [11–14], a highly representative SSL approach, has achieved further success in the field of natural images. These methods define a contrast prediction task that maximizes the similarity

* Corresponding author.
E-mail addresses: hj1992884@163.com (J. Huang), yangchuan05@126.com (C. Yang), 149265005@qq.com (S. Liu).

https://doi.org/10.1016/j.compbiomed.2023.107337
Received 29 April 2023; Received in revised form 13 July 2023; Accepted 7 August 2023
Available online 12 August 2023
0010-4825/

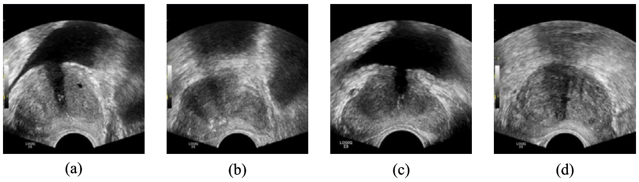

Fig. 1. Transrectal ultrasound images of the prostate. Fig. 1(a), (b) are normal prostate TRUS images; Fig. 1(c), (d) are prostate cancer TRUS images. It can be seen that the ultrasound images have problems such as noise, artifacts and intensity inhomogeneity.

of representations from different enhanced views of the same image, while maximizing the relative distance between representations from different images, by means of a contrast loss function.

However, it is relatively difficult to apply CL in the field of medical ultrasound images for several reasons. First, there exists a domain gap between natural and medical images [15]. Additionally, a training batch may contain multiple images of the same category, which means images of the same class will be treated as negative. This will impair the ability of the contrast loss to cluster similar images together, as different image classes produce higher loss values. Moreover, most standard CL-based algorithms mainly focus on the global features of the images, ignoring the influence of the relationships between feature maps on the construction of global representations. There are also numerous problems with our dataset of transrectal ultrasound prostate images (Fig. 1): the image resolution is low, there is a large amount of information jamming in the images, the high similarity between positive and negative samples, and the details of pathological structures cannot be displayed clearly. Thus, it is difficult to learn its key diagnostic information using the CL algorithm directly. Furthermore, the encoder and decoder architecture adopted by CL algorithm leads to redundant information by repeatedly extracting similar low-dimensional features in the image, while struggling to model high-dimensional features constructed by the projection layer, which causes the model to construct non-optimal feature representations for different correlation classes.

To address the aforementioned problems, we propose a self-attentive convolutional neural network, called **S**elf-adaptive **D**ual **A**ttention **B**ootstrap **L**earning (SDABL), based on contrast learning. To solve the first problem, we use unlabeled raw datasets for training, and our approach discards the construction of negative sample pairs and only uses loss calculations that approximate similar positive sample pairs to guide the model representation embedding. For the second problem, this paper introduces a new spatial attention module named Self-Position Attention Module (SPAM) to enhance the network's learning of spatial location information. Compared with the traditional spatial attention module, SPAM discards the convolution operation when sampling and focuses on the spatially valid information of its own feature map. We also propose an adaptive maximum channel attention module (CAAM), which guides the construction of its own channel feature map. CAAM has a small number of parameters and more nonlinearities to improve the accuracy of network-generated feature map channel weights. We embed the two proposed attention modules into the underlying CNN to perform feature graph information optimization. This network maps the input TRUS image to the embedded feature space, where we use the minimum contrast loss to maximize the similarity between different feature views of the same ultrasound image. The main contributions of this work are as follows:

(1) We propose an adaptive contrast learning model, SDABL, for prostate cancer screening. Our model can learn valid visual representations without relying on annotated data, alleviating the reliance on large-scale annotated training sets and allowing the development of pre-trained models specific to medical imaging.
(2) We developed a self attention bootstrap method to maximize effective representation. We proposed SPAM and CAAM modules are complementary, and allow us to construct effective feature maps based on both local and global information.
(3) We conduct a comprehensive experiment in which we compare the proposed SDABL network with the mainstream CL algorithm on clinical prostate TRUS data. The final experimental results demonstrated the superiority of our method over the mainstream algorithm.

The rest of the paper is organized as follows. The second section reviews research related to automatic prostate cancer diagnosis in deep learning, visual representation and contrast learning, and self-attentive mechanisms. The proposed approach that we present is in Section 3. Section 4 discusses the experiments and shows the visualization results. Finally, we conclude our report in Section 5.

2. Related work

In this section, we review recent developments in deep learning-based prostate cancer, applications of visual representation learning, and methods related to self-attention mechanisms.

2.1. Automatic prostate cancer diagnosis in deep learning

In recent years, the role of deep learning in prostate assisted diagnosis has rapidly expanded due to innovations in the field of computer vision. After training on large datasets, convolutional neural networks (CNNs) can achieve similar accuracy as imaging physicians with higher diagnostic efficiency [16]. Deep learning methods can help conserve precious medical resources and reduce the burden for imaging physicians. For the blurred pixel problem in MRI images, Ari et al. [17] proposed the use of a denoising filter with increasing the size of the dataset. They used a two-dimensional convolution neural network model, which produced high classification accuracy for prostate cancer MRI images. Ye et al. [18] proposed a two-stage diagnostic method for prostate tumors, which used two CNNs for feature extraction and cancer diagnosis. This method had an accuracy of 87.95% for prostate tumor determination as well as superior diagnostic speed. Duran-Lopez et al. [19] proposed a model of CNNs based on custom width and depth to solve the problem of rapid diagnosis of high-resolution prostate pathograms. They extracted and aggregated patch slices with diagnostic information from pathological maps, then performed sliding classifications using a patch puzzle.

In the field of biomedical models, data scarcity is a common limiting factor for technological development. Currently, there are some approaches which adopt transfer learning or generative model to alleviate this problem. Lu et al. [20] proposed a super-resolution reconstruction method based on ultrasound images of the prostate to improve the poor performance of classification networks with limited prostate data. The prostate images are first reconstructed, and then combined with the original dataset to improve the diagnosis of ultrasound image assisted classification of prostate cancer. Koc et al. [21] train VGG16 and VGG19 to extract different depth features in prostate MRI images using ImageNet pre-training. They combine these models with NCA analysis then use the kNN algorithm to classify the screening features. Yang et al. [22] proposed a weakly multi-task supervised framework which integrates different levels of information in the prostate cancer screening process. This research also addresses the problem that it is difficult to obtain accurately labeled Mp-MRI images in clinical practice.

Previous research has shown that deep learning networks can intelligently and efficiently diagnose prostate cancer. However, most of these methods are built on large scale labeled datasets, and also do not exploit the relationship between feature maps. Thus, these methods require constant acquisition of new datasets for retraining to be applied in medical scenarios.

2.2. Visual feature learning based on contrastive learning

One general solution to reduce reliance on manual annotation is to propose diverse pretexts for the network to solve. This process is called representation learning, also known as self-supervised learning [23]. In recent SSL methods, paradigms based on Contrastive Learning (CL) have shown great potential [24,25]. The core idea of CL is to construct sample diversity through data augmentation. These methods use the loss function to construct similar (positive) pairs of samples with close distances in projected high-dimensional space, and different (negative) pairs with relatively distant distances.

Since Wu et al. [24] proposed to use a memory bank to memorize representation vectors, there has been progress in comparison learning based on positive and negative sample pairs. Both SimCLR [11] and MoCo [12] used positive and negative sample pairs, while BYOL [26] abandoned this design in favor of the Siamese path paradigm for image representation learning. This paradigm guides the online path to predict different enhanced view representations of the same image generated by the target path, removing the need for the comparison prototype. There is a critical problem in the current CL algorithm research: it is difficult to maximize the original information of the image while removing irrelevant and redundant features in the process of self-representation construction. Xie et al. [27] addressed on this issue earlier. Their proposed representation learning method focuses on the consistency of local pixels, which is more suitable for pixel-level tasks such as segmentation. Huang et al. [28] pointed out that in the process of using projection layers, current SSL methods simply employ a uniform aggregation of pixels for embedding, while using different augmentations that may involve irrelevant interference and spatial misalignment. They proposed the LEWEL model to address this issue by utilizing adaptive aggregation of feature space information, thereby achieving more accurate alignment embeddings. Peng et al. [29] considered the effect of random sampling, an augmentation operation, on view quality. They proposed the method of training a preheated Grad-CAM [30] to locate the ROI region, and then perform center-compression sampling within the localized region. Following in their footsteps, we consider introducing more precise calculations in the characterization extraction algorithm by paying more attention to the potential semantic representation of the image.

In recent years, some studies have explored the application of SSL in medical image analysis. Chartsias et al. [31] proposed a supervised contrast learning approach to solve the bottleneck of data labeling for the unbalanced cardiac ultrasound data classification problem. Their approach achieved excellent performance for view classification with a small amount of labeled data. To address the issue of overfitting in deep classification models for medical images, Xing et al. [32] utilized the concept of contrastive learning and developed a contrastive knowledge distillation algorithm that preserves the classification relationship. By bringing closer the image pairs from the same category and pushing apart image pairs from different categories in both the teacher and student models, combined with a classification relationship preservation loss, they extracted the relational knowledge of the teacher model. Huang et al. [33] addressed the issue of automatic grading of diabetic retinopathy using two-part of datasets. First, they trained the target detection network to extract pathological plaque based on IDRiD dataset [34]. Then, they applied reasoning based on EyePACS dataset [35] to generate a high confidence lesion plaque image, and finally, combined it with CL algorithm to characterize DR grading with high discrimination by lesion plaque learning. In addition they investigated the impact of different data augmentation on the contrast learning algorithm.

Zhou et al. [36] argued that relying solely on contrastive estimation is not entirely optimal and that explicit solutions should be introduced to retain more information, following the design principles of SSL to maximize information preservation. They proposed Preservation Learning, which aims to enhance the performance of SSL models for medical image analysis by reconstructing diverse contextual information from the images. Dong et al. [37] proposed a framework for the first time that incorporates federated learning and comparative learning in order to efficiently utilize decentralized unlabeled medical data. The model can provide good representations for downstream tasks and achieves excellent performance in scenarios where the amount of labeled data is drastically reduced. Fernandez-Quilez et al. [38] used SimCLR as a basis for training prostate MRI images and used the resulting initialized model to achieve a classified diagnosis of PCa. This method employs the SimCLR framework to obtain a quality initialization model suitable for prostate MRI images directly. In contrast, our proposed method considers more about the applicability of the algorithm and is based on the fact that feature extraction is more difficult in TRUS images, which are poorly recognized by directly utilizing the contrast learning algorithm.

In the prior medical application scenarios, most of the comparison models include positive and negative sample pairs. This can lead to errors in the loss calculation. Furthermore, these methods may extract irrelevant and redundant feature information while embedding the representation alignment.

2.3. Attentional mechanisms in medical images

Attentional mechanisms have been widely used in computer vision tasks. Current methods can be broadly divided into two categories: positional attention and channeled attention, This plug-and-play approach has yielded quite promising results [39–41]. Sinha et al. [42] addresses the information redundancy problem posed by encoder–decoder architectures by capturing rich contextual dependencies through a guided self-attention mechanism. Attention mechanisms help capture detailed information about lesions in medical images. Yuan et al. [43] embedded an adaptive channel attention module in the U-Net [44] model, which automatic sorted important feature channels, integrated features at different levels using multi-level attention modules, and refinined the features of each individual layer through a self-attentive mechanism. The method achieves good retinal vessel segmentation performance with low model complexity. Sun et al. [45] addressed the problems of low contrast, speckle noise and low resolution of ultrasound images, through designed soft shape supervision, which used the cross-path attention mechanism. This method improved the performance of the model in detecting and segmenting the edge details of ultrasound images. Wu et al. [46] explored the use of Transformer combined with channel dimensional interaction. They improve the segmentation performance of the model by adding cross-dimensional interaction information to the feature map before the self-focus calculation. In order to segment clear sub-structures of the heart in computed tomography (CT) images, Park et al. [47] proposed a shape-aware attention module to guide the model to focus on the edge details between substructures. To assist in the diagnosis of pneumonia, Feng et al. [48] proposed condense attention module (CDSE) and the Multi-Convolutional Spatial Attention Module (MCSA), which are used to obtain channel weights and remove redundant information from feature maps, respectively. The method leverages the internal potential relationships of the feature map and overcomes the limitations of the traditional attention mechanism. To summarize, it is feasible for us to embed channel attention and spatial attention in a CNN to improve classification of TRUS images.

3. Methods

In this subsection, we first introduce our SDABL network in Section 3.1. Then, we describe our self-attention module in Section 3.2, finally in Section 3.3 we show our dual-path pipeline and our global consistency loss function.

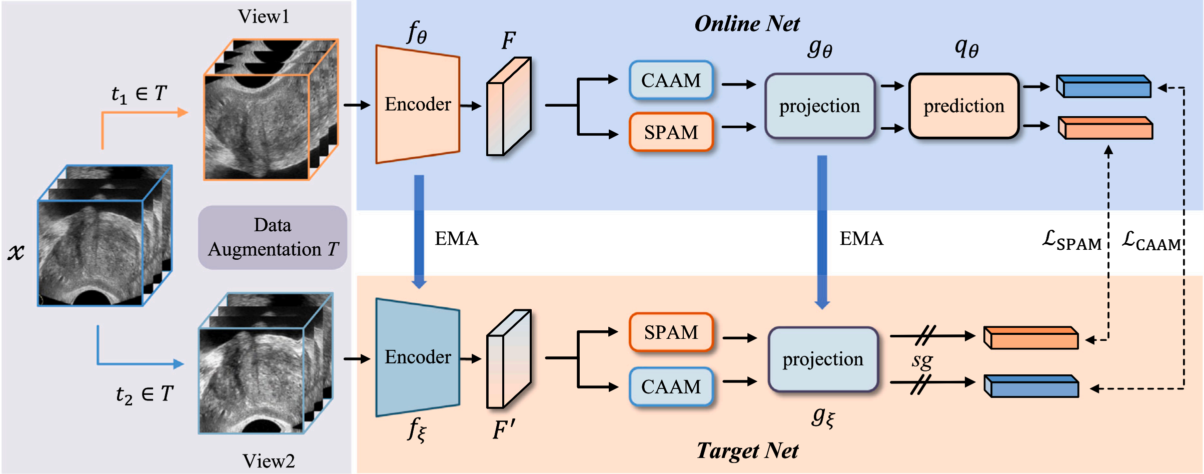

Fig. 2. The Self-supervised Dual-head Attentional Bootstrap Learning Network (SDABL) architecture proposed in this paper. The input x represents an image of a batch size, The t_1 data augmentation used a vertical flip and a random crop, t_2 data augmentation used color dithering and random crop operations, EMA means Exponential Moving Average, The symbol // means to stop the gradient operation. Our SDABL model uses global consistency loss to maximize the consistency between embedding features, At the end of training, everything except f_θ is discarded and F is used as the image representation.

3.1. Overall architecture

The SDABL model aims to learn an effective self-representation in unlabeled complex medical images using CL methods. In this work, the framework of the proposed SDABL is shown in Fig. 2, which can be divided into a data augmentation part, a dual-path network, and a loss calculation part.

For the input, assume that $S = \{x_i; i = 1, \ldots, N\}$ denotes a randomly sampled batch of images of size N. For each of these images x_i, we use two sets of randomly combined augmentation operators $\left(t_1, t_2 \in T\right)$. We produce the augmented View as follows: $View_1 = t_1\left(x_i\right)$ and $View_2 = t_2\left(x_i\right)$, The specific transforms used in data augmentation T are described in detail in Section 4.1.2. The two augmented views are then sent to the Online-Net and Target-Net dual-path pipelines for feature extraction and embedding alignment.

The dual-path network consists of a pair of encoders f_θ and f_ξ with shared parameters, a pair of decoders g_θ and g_ξ with shared parameters, and a separately set decoder q_θ in Online-Net. We only update the parameter θ in Online-Net, and perform a Exponential Moving Average (EMA) update in Target-Net. After passing the attention module and aligning the representation embeddings, we use global position loss jointly with global channel loss to minimize the consistency between features. In the final linear inference stage, we only keep the encoder f_θ and evaluate it linearly in the Online-Net path.

3.2. Dual self-attention block

In the dual-head self-attention module, we incorporated two cross-path attention mechanisms. This mechanism sets the feature maps of the two paths as Query (Q) and Key (K) for interaction, resulting in the generation of two weighted feature maps. After receiving two augmented Views $View_1$, $View_2$, as input to the encoder, we decrease the height and width and increase the channel size in the feature map $F, F' \in R^{C\times H\times W}$, where C, H, W denote the number of channels, height and width of the feature map, respectively. We split F and F' into two inputs, for the SPAM module and CAAM module respectively.

The SPAM and CAAM module schematics are shown in Fig. 3. In SDABL we inserted two groups of this module in parallel. Specifically, in SPAM we obtain Q and K inputs for image information interaction processing by mapping feature maps. SPAM first flattens and reconstructs the feature map $F \in R^{C\times H\times W}$ to obtain $F_1 \in R^{C\times(H\times W)}$ and $F_2 \in R^{(C\times H)\times W}$. We compute spatial self-attentive weights through matrix multiplication and the softmax activation function. We then multiply it with the reshape form of the feature map F to get our spatial attention interaction feature map. Finally, we obtain our output features by shape recovery and then add it to our original input. The calculation is expressed by the following equation:

$$Self\text{-}Attention(SPAM) = R'\left\{R(K)\otimes Softmax(R^T(Q)\otimes R(Q))\right\}\oplus K \tag{1}$$

where R denotes the reshape operation, R^T is the transpose after reshape operation, R' is the recovery vector is tensor shape, $\otimes$ denotes a matrix multiplication operation and $\oplus$ denotes a matrix addition operation. The SPAM module is responsible for learning the spatial location relationship of feature maps, which indicate the importance of the location of each spatial feature. Similarly, for the CAAM module, direct mapping of the feature map yields Q and K (where $Q \triangleq K$). We then transpose and reconstruct the feature map $F \in R^{C\times H\times W}$ to obtain $F_1' \in R^{C\times(H\times W)}$ and $F_2' \in R^{(C\times H)\times W}$. Through matrix multiplication and activation function by softmax, we compute the channel self-attention weight, then multiply with the reshape form of the feature map F and recover the original shape. Finally, our global self-channel feature map is obtained by adaptive maximum pooling, and we obtain the output features by matrix addition with the original feature map. The CAAM is calculated by the following equation:

$$Self\text{-}Attention(CAAM) = AvgMaxpool\left[R'\{Softmax(R(Q)\otimes R^T(Q))\otimes R(K)\}\right]\oplus K \tag{2}$$

Our CAAM module is responsible for extracting the global channel information. We use adaptive maximum pooling operation after obtaining the channel attention feature map. This operation refines the effective information of the feature map, enhances the influence of the surrounding local features, and removes part of the redundant information of the feature map.

Our approach here is different from most attentional mechanisms [45,49]. In our self-attentive computation process, we discard the convolution operation in order to focus on the original encoding information of the feature map. We use this design to bring more self-focused spatial invariance to subsequent feature embeddings. Furthermore, we employed parallel additive computation, which involves adding a skip connection in the original self-attention path. This enables us to highlight the essential parts of the feature maps by adding the attention encoding with the original feature maps. Because the self-focus mechanism transforms the input feature map from two-dimensional information to one-dimensional information, the spatial

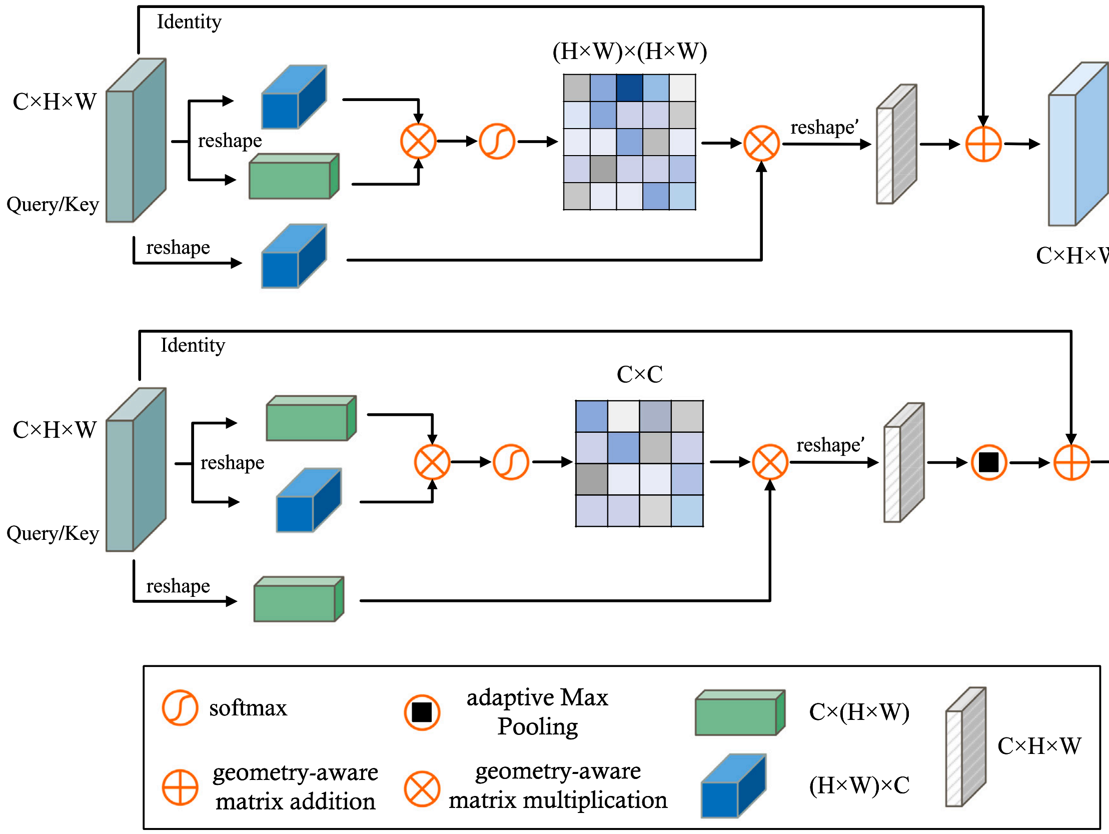

Fig. 3. Our SPAM module (top) and CAAM module (bottom). "*reshape*" operation means that the feature maps are flattened into vectors, "*reshape′*" operation means recover the vector as the feature map tensor shape. Green cubes indicate direct use of reshape operation, and the blue cube indicates the use of reshape followed by the transposition operation.

location information of the feature map is lost. The advantage of SPAM module and CAAM module is the maximum preservation of spatial and inter-channel information. We aggregate the two attention modules, SPAM and CAAM, in parallel, to complementarily mine the relationships between feature channels and between location spaces. The experimental results demonstrate that this method significantly improves the performance of prostate TRUS image recognition.

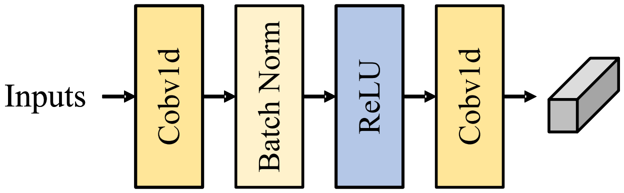

Fig. 4. Our adopted Conv1d combined BN-ReLU embedding alignment.

3.3. Dual-path network

Encoder. The backbone network used in this study is the ResNet [50] series. To make better use of the original image information while considering global channel information and spatial location information, we drop the adaptive average pooling layer and the fully-connected layer from the original ResNet network.

Embedding space. The prior two-path Siamese comparison model [11,12,26], use the MLP to construct both its projection layer and prediction embedding layer. This mapping approach is an embedding of global features in uniform aggregation, lack of processing to spatial information. As shown in Fig. 4 our Projection and Prediction use the Conv1d-BN-ReLU-Conv1d structure. Suppose the input channel dimension is C and the final number of embedded channels is set to D. We then denote the dimensional change of the projection and embedding as $g_\theta : (C \rightarrow 2C \rightarrow D)$, $q_\theta : (D \rightarrow C \rightarrow D)$, $q_\xi : (D \rightarrow 2C \rightarrow D)$. We analyze the impact of the number of channels in the predicted embedding space D, on the downstream classification performance in detail in Section 4.6.2.

The goal of SDABL is to learn good representation of F, then transfer to downstream tasks. The Online-Net is defined by a set of weights θ. The Target-Net has the same structure as the Online-Net and uses a separate set of weights ξ. According to our architecture (Fig. 2), the two Augmented Views $View_1$ and $View_2$ are processed by the encoders f_θ and f_ξ, respectively. Then, with our self-attention module, we obtain the feature map of the relationship between attention location encoding and channels. Finally the representation vector is obtained by aligning our embedding. We will extract the embedding feature vector as $P_1(\cdot) \triangleq q_\theta(g_\theta(F_1(\cdot)))$ and $Z_2(\cdot) \triangleq stopgrad(g_\xi(F_2'(\cdot)))$, where $F_1(\cdot)$ denotes the input $View_1$ features extracted by our SPAM module and CAAM module, $P_1(\cdot)$ denotes the corresponding obtained two sets of embedding tensor. $F_2'(\cdot)$ denotes the two sets of features extracted from $View_2$ into our SPAM module and CAAM module, $Z_2(\cdot)$ denotes the corresponding obtained two sets of embedding vectors. Then we set $P_1(\cdot)$ as the embedding prediction vector and $Z_2(\cdot)$ as the target vector, measure the distance between the two in the embedding space. We define the regularized cosine similarity minimization:

$$\mathcal{D}(P_1(\cdot), Z_2(\cdot)) = \frac{P_1(\cdot)}{\|P_1(\cdot)\|_2} \cdot \frac{Z_2(\cdot)}{\|Z_2(\cdot)\|_2} \tag{3}$$

Where $\|\cdot\|_2$ denotes L2 regularization (l_2-norm), Eq. (1) is defined for each image. The total loss is averaged over all images, and its minimum possible value is 0. We swap $View_1$ and $View_2$ and input them into our dual path network. We denote as $P_2(\cdot) \triangleq q_\theta(g_\theta(F_2(\cdot)))$ and

Table 1
Distribution of our TRUS dataset.

TRUS data	Number of patients	Number of images
Malignant	74	517
Benignant	110	678
Total	184	1195

$Z_1(\cdot) \triangleq stopgrad(g_\xi(F'_1(\cdot)))$, and obtain the cosine similarity between $P_2(\cdot)$ and $Z_1(\cdot)$ by:

$$\mathcal{D}(P_2(\cdot), Z_1(\cdot)) = \frac{P_2(\cdot)}{\|P_2(\cdot)\|_2} \cdot \frac{Z_1(\cdot)}{\|Z_1(\cdot)\|_2} \tag{4}$$

The stopgrad operation (Fig. 2) was used to prevent the model training from collapsing.

We construct the Online-Net and Target-Net in a form similar to $P(\cdot)$ and $Z(\cdot)$ setting each other as pseudo-labels to guide the network for learning. It is important to note that Target-Net does not perform a gradient backward update in the training step. Following the settings in [26], Target-Net weights ξ are updated only by the exponential moving average (EMA) of the weights θ in the corresponding Online-Net, as shown below

$$\xi = (1 - \lambda) * \theta + \lambda * \xi \tag{5}$$

where λ denotes the target decay rate.

The similarity calculation is the key to the loss function calculation for comparison learning. We extend the original loss calculation of [26] with the proposed joint minimization loss definition objective. As shown in Eqs. (1)(2), we obtained the cosine similarity of the output tensor. The function $\mathcal{D}$ considers the global consistency between different augmented views by minimizing the difference between the normalized Online-Net features and the Target-Net features. Specifically, we propose a hyperparameter β to fuse the two attentional features. Thus, we make network training focus on both consistency of location details and consistency of global channels and guide the network to retain more effective structure of the characterization parameters when updating. We define the loss component focusing on location information and the loss focusing on global channel information as:

$$\mathcal{L}_{SPAM} = \mathcal{D}(q_\theta(g_\theta(F(SPAM))), g_\xi(F'(SPAM))) \tag{6}$$

$$\mathcal{L}_{CAAM} = \mathcal{D}(q_\theta(g_\theta(F(CAAM))), g_\xi(F'(CAAM))) \tag{7}$$

$$\mathcal{L} = (1 - \beta) * \mathcal{L}_{SPAM} + \beta * \mathcal{L}_{CAAM} \tag{8}$$

As shown in Fig. 2, we obtained $\mathcal{L}_{SPAM}$ loss and $\mathcal{L}_{CAAM}$ loss separately. The loss function is then calculated directly using linear fusion. Finally, according to [26] and [13], we set the objective loss function as:

$$\mathcal{L}_{Total} = 0.5 * \mathcal{L}_{(1,2)} + 0.5 * \mathcal{L}_{(2,1)} \tag{9}$$

In Eq. (9), $\mathcal{L}_{(1,2)}$ means $View_1$ input Online pipeline and $View_2$ input Target pipeline. $\mathcal{L}_{(2,1)}$ means $View_2$ input Online pipeline as the prediction value and $View_1$ input Target pipeline as the label value. The pseudo-code of our SDABL algorithm is visible in Algorithm 1.

4. Experiments and discussions

In this section, we first present the experimental pre-training setup, which includes the sampling process of the dataset, the data augmentation strategy and the optimization strategy and parameters, and then introduce commonly used classification evaluation metrics. Finally we analyze in detail the experimental results on the prostate TRUS dataset and demonstrate the validity of the proposed attention mechanism through visualizations.

4.1. Pre-training settings

4.1.1. Dataset

The prostate TRUS data used in this study was collected from patients attending the ultrasound department of the First Affiliated Hospital of Jinan University from May 2018 to March 2022. These images were collected by the patient during a transrectal ultrasound-guided, puncture biopsy procedure of the prostate. Puncture biopsy applies the Logiq E9 ultrasound diagnostic instrument from GE, USA, ultrasound guidance with the aid of the IC5-9D intracavitary probe with a frequency of 3 to 10 MHZ. Currently, we have collected a total of 1195 images in the original dataset. These were collected from 110 positive and 74 negative patients at different periods (See Table 1.)

4.1.2. Data augmentation

Due to the inclusion of hospital information labels and acquisition time stamps in the collected TRUS images, we employ center cropping to extract the rectangular region of the prostate as the ROI. Subsequently, we apply data augmentation operations to the cropped ROI region. Following the procedures outlined in [26], we set up the data augmentation module in Fig. 2. Two random transformations are performed on the input image, generating two views corresponding to the two paths:

$$View_1 = t_1(x), View_2 = t_2(x) \tag{10}$$

For this study, we used the following augmentation operations:

(1) Color jitter. The application probability is 80%. Dithering Brightness is set to 0.7, Contrast is set to 0.7, Saturation is set to 0.7, and Hue is set to 0.2.

(2) Convert to grayscale. Probability of application is 20%.

(3) Horizontal Flip. Probability of application is 50%.

(4) Vertical Flip. Probability of application is 50%.

(5) Random cropping and scaling. We crop randomly from the original image with a size ranging from 85% to 100% of the input image size. To ensure that all images are processed in the subsequent work, we Resize both views to the same size (e.g. 256 × 256) at the end of the augmentation.

(6) Gaussian blur. Probability of application is 50%. Gaussian kernel σ sampled from a uniform distribution between $0.5 \sim 1$.

The six augmentation operations mentioned above are combined in a sequential manner. When using PyTorch functions to load the data, for each batch and each view within the batch, the corresponding combination of augmentation operations is applied randomly with probability.

4.2. Evaluation criteria

To comprehensively measure the effectiveness of the proposed network, we employed accuracy, recall, precision and F1 score as evaluation metrics. Accuracy measures overall network performance. Recall measures the proportion of true positive images among all TRUS images. Precision measures the fraction of true positive images among all predicted positive images. F1 score is the summed average of accuracy and recall. The formulas are defined as follows:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \tag{11}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{12}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{13}$$

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision + Recall}} \tag{14}$$

Where TP is true positive, TN is true negative, FP is false positive and FN is false negative.

Algorithm 1 The algorithm of SDABL, Pytorch-like

```
Input: unlabeled training data $S = \{x_i; i = 1, \dots, N\}$, batch size $N$, Encoder network: $f_\theta, f_\xi$,
pre-train projection head: $g_\theta$, pre-train prediction head: $g_\theta, g_\xi$, Augmentation: $aug_1, aug_2$,
SPAM module: spam, CAAM module: caam, Hyperparameters: weight $\beta$, stopgrad: sg
for $x$ in loader:
    $\mathrm{View}_1, \mathrm{View}_2 = aug_1(x), aug_2(x)$ # random augmentation
    Input View$_1$ into the Online-Net, View$_2$ into the Target-Net, compute:
    get: $p_1 \triangleq q_\theta(g_\theta(spam(f_\theta(View_1))))$, $p_1' \triangleq q_\theta(g_\theta(spam(f_\theta(View_1))))$
    and $z_2 \triangleq sg(g_\xi(spam(f_\xi(View_2))))$, $z_2' \triangleq sg(g_\xi(spam(f_\xi(View_2))))$
    Input View$_2$ into the Online-Net, View$_1$ into the Target-Net, compute:
    get: $p_2 \triangleq q_\theta(g_\theta(spam(f_\theta(View_2))))$, $p_2' \triangleq q_\theta(g_\theta(spam(f_\theta(View_2))))$
    and $z_1 \triangleq sg(g_\xi(spam(f_\xi(View_1))))$, $z_1' \triangleq sg(g_\xi(spam(f_\xi(View_1))))$
    Calculation of similarity: $\mathcal{D}(p_1, z_2) = \frac{p_1}{\|P_1\|_2} \cdot \frac{z_2}{\|z_2\|_2}$, and similarly get: $\mathcal{D}(p_1', z_2')$, $\mathcal{D}(p_2, z_1)$, $\mathcal{D}(p_1', z_2')$
    Calculation Total loss
    $\mathcal{L}_{(1,2)} = (1 - \beta) * \mathcal{D}(p_1, z_2) + \beta * \mathcal{D}(p_1', z_2')$
    $\mathcal{L}_{(2,1)} = (1 - \beta) * \mathcal{D}(p_2, z_1) + \beta * \mathcal{D}(p_2', z_1')$
    $\mathcal{L}_{(\mathrm{Total})} = 0.5 * \mathcal{L}_{(1,2)} + 0.5 * \mathcal{L}_{(2,1)}$
    Back-propagate
    $\mathcal{L}_{(\mathrm{Total})}.backbone$
    SGD update
    $update(\theta)$
end
Return: encoder network $f_\theta$
```

Table 2
Comparison with mainstream contrastive learning methods.

Methods	Malignant			Benignant			Acc. [%]
	F1	Recall	Precision	F1	Recall	Precision	Test
Doctor	56.50	42.10	85.80	**68.10**	**90.90**	54.50	63.20
MoCo [12]	76.97	77.22	76.73	62.56	62.24	62.89	73.43
	77.95	81.65	74.57	59.67	55.10	65.06	73.82
	79.88	84.18	76.00	62.57	57.14	69.14	74.89
BYOL [26]	77.06	80.25	74.12	59.46	55.56	63.95	73.04
	77.88	79.62	76.22	62.83	60.61	65.22	73.42
	78.44	83.44	74.01	59.55	53.54	67.09	73.43
SimSiam [13]	80.11	89.24	72.68	56.25	45.92	72.58	73.46
	77.97	87.34	70.41	50.63	40.82	66.67	70.70
	79.55	89.87	71.36	52.90	41.84	71.93	72.26
SimCLR [11]	68.00	75.89	61.59	59.60	52.68	68.60	67.85
	69.02	78.57	61.54	59.07	50.89	70.37	66.07
	71.07	76.79	66.15	66.02	60.71	72.34	71.87
SDABL (D = 256)	80.36	84.71	76.44	64.09	58.59	70.73	77.73
	81.27	**89.81**	74.21	60.61	50.51	75.76	77.34
	80.12	82.17	**78.18**	66.32	63.64	69.23	**78.90**

The backbone networks used in the order from top to bottom correspond to ResNet-18, ResNet-34, and ResNet-50, respectively. The basic CL models in this context utilize MLP for projection embedding. Our SDABL setting $\beta = 0.5$.

Table 3
Comparison of different self-attentive modules embedded with ResNet-50 as Backbone.

Methods	Malignant			Benignant			Acc. [%]
	F1	Recall	Precision	F1	Recall	Precision	Test
Baseline	77.64	76.22	**79.11**	62.11	**64.13**	60.20	75.05
Dual Net [49]	78.86	**86.79**	72.25	54.32	45.36	67.69	73.43
PAM [49]	66.38	69.64	63.41	62.91	59.82	66.34	70.08
PPM [51]	78.05	81.01	75.29	60.87	57.14	65.12	74.60
PPM+	79.13	80.38	77.91	64.92	63.27	66.67	74.60
SSSB [45]	77.65	83.54	72.53	55.81	48.98	64.86	72.65
SSSB+	78.61	86.08	72.34	55.42	46.94	67.65	74.21
SDABL	**82.67**	86.62	79.07	**68.85**	63.64	**75.00**	**80.46**

We set the embedding dimension $D = 256$ dimensions, the "+" symbol in the table indicates the use of skip connections, similar to our proposed attention modules, where the original input feature map is added to the feature map on which the attention operation is performed, where PPM is our SPAM module prototype.

4.3. Comparison with mainstream contrastive learning methods

We use ResNet-18, Resnet-34, and ResNet-50 as our backbones and unify the embedding dimension as $D = 256$. The base CL method used also removes the last two layers from the ResNet network. As seen in Table 2, for two sets of memory bank-based models with positive and negative samples (MoCo, SimCLR), the model fits better as the depth of backbone network increases. The performance of BYOL method, on the other hand, does not change significantly with the depth of the network. One of the SimSiam methods is a lightweight CL algorithm, and thus works best when ResNet-18 is used. Our SDABL method works best for linear fine-tuning verification when using ResNet-34 because it is built on the BYOL framework. We also find that the performance improvement of SDABL compared to the BYOL [26] model is more pronounced as the feature depth increases. The results of this experiment demonstrate the effectiveness of our self-attentive approach and the efficient extraction of deep semantic representations, resulting in improved recognition performance of the model. In contrast, physicians had the best indicators of accuracy for positive samples and recall for negative samples, but exhibited fairly poor performance overall.

4.4. Comparison with recent advanced attentional methods

This section discusses how the SDABL method compares with recent advanced attentional methods. Our baseline setup encoder uses the ResNet-50 network and removes the last two layers of the network and uses the nonlinear convolutional layer Conv1d-BN-ReLU-Conv1d instead of MLP. As shown by the results in Table 3, our dual path self-attention mechanism performs best in linear fine-tuning performance. One possible explanation for our method's superior performance is that the convolution operations used in the other self-attentive modules may destroy the representational consistency from image augmentation in the CL algorithm. While our module is more aligned using image raw information, by simply self-focusing for embedding representation of image features alignment is more complete. Additionally, in both the

Table 4
Ablation experiment of TRUS.

Ablations	Malignant			Benignant			Acc. [%]
	F1	Recall	Precision	F1	Recall	Precision	Test
BYOL [26]	78.93	84.71	73.89	59.43	52.53	68.42	73.43
baseline	77.64	76.22	**79.11**	62.11	**64.13**	60.20	75.05
baseline+SPAM	79.50	81.53	77.58	65.26	62.63	68.13	76.56
baseline+CAAM	82.32	85.99	78.95	68.48	63.64	74.12	78.51
SDABL(Ours)	**82.67**	**86.62**	79.07	**68.85**	63.64	**75.00**	**80.46**

Using ResNet-50 as the linear classification result of Backbone, we set the embedding dimension $D = 256$ and use the cosine similarity measure. Where baseline is based on ResNet-50 with the adaptive pooling and FC layers removed, and a Conv1d-BN-ReLU-Conv1d projection and embedding structure is used.

attention modules of SDABL, a skip connection pathway is incorporated. Its purpose is to ensure stable and efficient output of the SPAM and CAAM models regarding the feature maps. This allows information to flow between different layers, making it easier for the network to learn complex nonlinear function mappings.

4.5. Comparison with supervised learning methods

For the most recent supervised learning method, the literature [52] used 66 positive and 103 negative samples as a test set for TRUS classification. This method achieved a positive sample F1 score of 65% and a negative sample F1 score of 83%. Compared to this supervised model, our model has a higher identification rate for positive samples, but poorer discrimination for negative samples. The literature [20] directly uses supervised learning methods for TRUS classification, and obtains a combined accuracy of 84.3%. After acquiring the hyper-segmented resolution reconstructed TRUS images and combining them with the original dataset, its combined accuracy rises to 86.7%. From these results, we can see that the linear fine-tuning results of our SDABL model achieves comparable performance to supervised learning.

4.6. Ablation studies

Experimental details: In our ablation experiments, the image size of the input encoder is uniformly 256 × 256 pixels. We use a batch size of 32, the initial learning rate is 0.0001, the weight decay is 10^{-4}, the temperature term τ is 0.1, and the momentum parameter λ of the encoder is fixed to 0.99. We use stochastic gradient descent (SGD) with momentum of 0.9 to minimize our loss function and use cosine annealing optimization [53] to train the model for 100 epochs. In the downstream linear fine-tuning of the classification task the batch size is 32, the initial learning rate is 0.001, the weight decay is 10^{-3}, and the SGD with a momentum of 0.9 is used, used the cross-entropy loss function, and used the same cosine annealing optimization [53], and the fitting training period was also set to 100.

Experimental results: Table 4 reported the fine-tuned linear classification results under our TRUS dataset. To ensure a fair comparison, the experimental parameters in this study were set to be consistent. As shown in the Table above, baseline compared to the [26] model, abandoning the MLP uniform projection improved the overall accuracy by 1.62%. Model 3 introduces the SPAM attention module, which improves the accuracy, positive F1 score, and negative F1 score by 1.51%, 1.84%, 3.15%, respectively. Model 4 introduces the CAAM attention module into the baseline, which adaptively samples the feature maps of the channel dimensions, then concatenates the outputs and generates spatial masks by adaptive maximum pooling. Compared with model 3, the introduction of CAAM improves the accuracy, positive F1 score, and negative F1 score by 3.46%, 4.68%, and 6.37%, respectively. Clearly, the CAAM module improves upon the CL algorithm encoder by introducing more nonlinearities. The idea is to use the channel characteristics themselves to exclude redundant information. Model 5 is our proposed method SDABL, which inserts the SPAM module and CAAM module in parallel before the projection layer and after the Backbone layer.

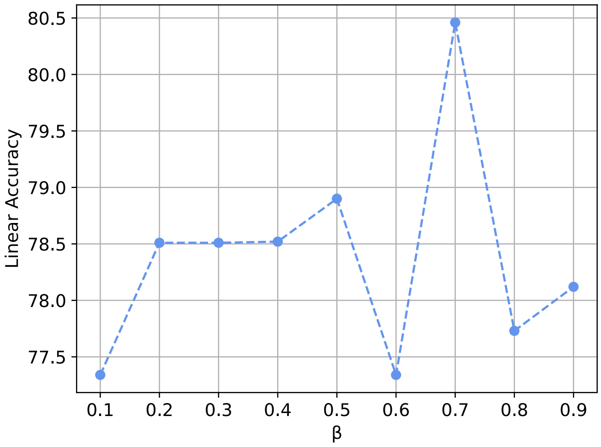

Fig. 5. Linear validation effect of SDABL with different β-Weight effects. We fixed the embedding dimension $D = 256$, used ResNet-50 as the backbone network, and the target loss was constructed using cosine similarity.

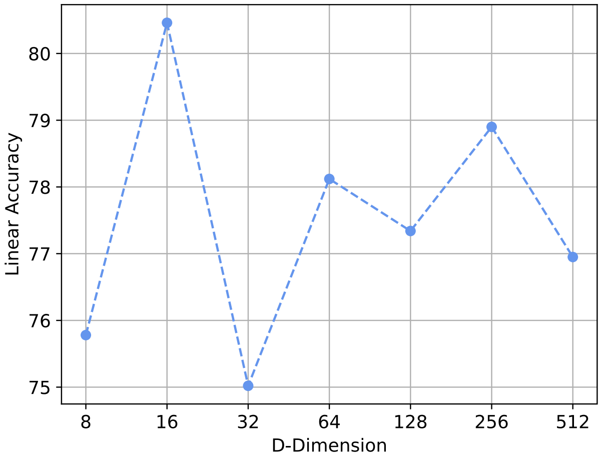

Fig. 6. Effect of different alignment embedding dimensions on linear classification of SDABL. We set the weights fixed at $\beta = 0.5$, use ResNet-50 as the Backbone, and the loss function is constructed using the cosine similarity.

The results show that the classification performance of SDABL validation is better than that of the SPAM alone or CAMM alone. The overall accuracy of SDABL was 80.46%, with a positive F1 score of 82.67% and a negative F1 score of 68.85%.

4.6.1. Influence of the loss weight

As shown in Fig. 5, we tested the linear classification effect of SDABL with different loss weights β. In order to investigate the extent to which the weight β affects the performance of SDABL, we observe that the model performs best when $\beta = 0.7$, at this time, more attention is paid to the feature map relationship between channels. Except for the case of $\beta = 0.6$, the model performs relatively poorly at $\beta \leqslant 0.5$, a comparable effect is produced when β is set to a larger value. The overall observation shows that our adaptive feature extraction operation and weighting scheme are effective. This method incorporates global loss and local loss to guide the model construction in favor of overall representation learning.

4.6.2. Influence of the dimensionality of the aligned embeddings

We also conducted experiments to investigate the effect of the final embedding space D on our model. As described in 3.3, we set the

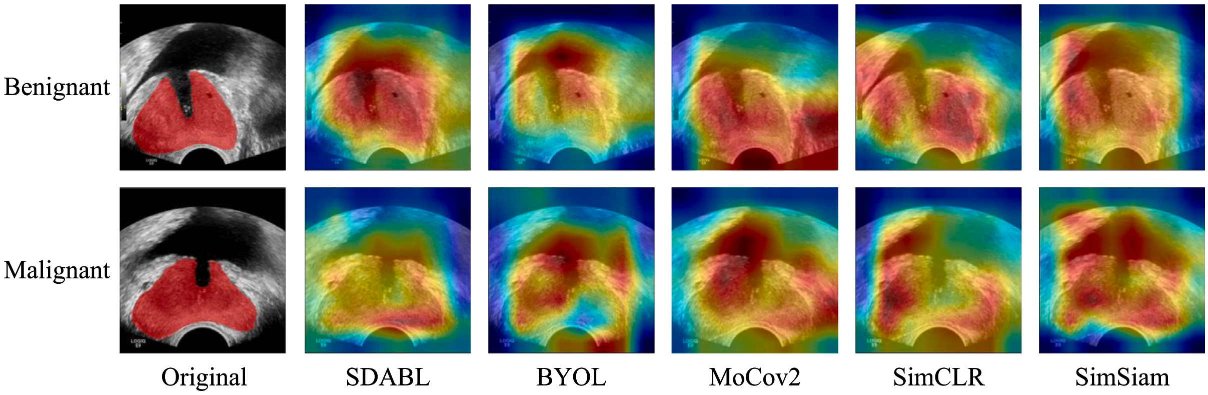

Fig. 7. Results of Grad-CAM visualization with different comparison learning algorithms.We use the mainstream CL model with Grad-CAM to visualize the last convolutional layer. Column 1 represents the original prostate ultrasound image, column 2 is our SDABL method, and the rest is the CL algorithm obtained using the ImageNet pre-trained model.

inputs to the projection and prediction heads to be fixed to C. The experimental results are shown in Fig. 6, from the results we can see that the model classification performance reaches its maximum at $D = 16$, excluding that the accuracy of the model increases with the increase of dimension D. Considering the suitability of our dataset for the CL method, it is possible that the small number of data categories in linear classification caused anomalous results when embedding in low dimensions. Overall, the expansion of the alignment embedding requires a greater amount of effective information to learn the overall representation. However, when the dimension is too large ($D > 256$), the extra information may be redundant and thus the performance does not improve further.

4.7. Visualization

To prove the validity of our method, we used the Grad-CAM [30] method for visualization-assisted verification. As shown in Fig. 7, most contrast learning algorithms focus on the overall region of the prostate image, which implies that these algorithms are more concerned with the global consistency of the image. In contrast, the visualization for our SDABL model shows that the positive image samples focus less on the prostate artifacts, while the negative samples focus more on the prostate area. Thus, our proposed self-attention method visually outperform other CL methods.

5. Conclusion

In this work, we proposed the self-supervised dual-head attentional bootstrap learning network, a new contrast learning model for the preliminary diagnosis of prostate cancer. This network incorporates two attention mechanisms, the SPAM module and the CAAM module. The SPAM module aids the extraction of self-position attention information, while the CAAM module is responsible for global effective representation selection. It encourages our network to learn more about the real diagnostic features from the prostate ultrasound images and provides an excellent pre-trained model for downstream classification tasks. In addition, we also conducted a detailed test of the parameters of the CL algorithm. This is specifically done by connecting a linear classification layer after the encoder. We show that our method has superior classification performance on TRUS images in the limited labeling case. For future work, we will investigate the self-supervised detection algorithm to localize the ROI region of prostate cancer in complex backgrounds and without labels. This will improve the accuracy and reliability of intelligent aided diagnosis of prostate ultrasound images and promote the application of computer vision technology in medical intelligent aided diagnosis.

Declaration of competing interest

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 in part by the National Natural Science Foundation of China (62176067); Scientific and Technological Planning Project of Guangzhou (201903010041, 202103000040); Key Project of Guangdong Province Basic Research Foundation (2020B1515120095); Project Supported by Guangdong Province Universities and Colleges Pearl River Scholar Funded Scheme (2019); Industry-University-Research Innovation Fund for Chinese Universities (2023HT023).

References

[1] Changfa Xia, Xuesi Dong, He Li, Maomao Cao, Dianqin Sun, Siyi He, Fan Yang, Xinxin Yan, Shaoli Zhang, Ni Li, et al., Cancer statistics in China and United States, 2022: Profiles, trends, and determinants, Chin. Med. J. 135 (05) (2022) 584–590.

[2] Nigel Hawkes, Cancer survival data emphasise importance of early diagnosis, 2019.

[3] David Eldred-Evans, Hashim U. Ahmed, Population-based prostate cancer screening with magnetic resonance imaging or ultrasonography—The IP1-prostagram study—Reply, JAMA Oncol. 7 (10) (2021) 1575–1576.

[4] Hoo-Chang Shin, Holger R Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel Mollura, Ronald M Summers, Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning, IEEE Trans. Med. Imaging 35 (5) (2016) 1285–1298.

[5] Xu Lu, Shulian Zhang, Zhiyong Liu, Shaopeng Liu, Jun Huang, Guoquan Kong, Mingzhu Li, Yinying Liang, Yunneng Cui, Chuan Yang, et al., Ultrasonographic pathological grading of prostate cancer using automatic region-based gleason grading network, Comput. Med. Imaging Graph. 102 (2022) 102125.

[6] Xinggang Wang, Xianbo Deng, Qing Fu, Qiang Zhou, Jiapei Feng, Hui Ma, Wenyu Liu, Chuansheng Zheng, A weakly-supervised framework for COVID-19 classification and lesion localization from chest CT, IEEE Trans. Med. Imaging 39 (8) (2020) 2615–2625.

[7] Mehdi Noroozi, Paolo Favaro, Unsupervised learning of visual representations by solving jigsaw puzzles, in: European Conference on Computer Vision, Springer, 2016, pp. 69–84.

[8] Junnan Li, Pan Zhou, Caiming Xiong, Steven C.H. Hoi, Prototypical contrastive learning of unsupervised representations, 2020, arXiv preprint arXiv:2005.04966.

[9] Yuanfan Guo, Minghao Xu, Jiawen Li, Bingbing Ni, Xuanyu Zhu, Zhenbang Sun, Yi Xu, HCSC: Hierarchical contrastive selective coding, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 9706–9715.

[10] Chen Feng, Ioannis Patras, MaskCon: Masked contrastive learning for coarse-labelled dataset, 2023, arXiv preprint arXiv:2303.12756.

[11] Ting Chen, Simon Kornblith, Mohammad Norouzi, Geoffrey Hinton, A simple framework for contrastive learning of visual representations, in: International Conference on Machine Learning, PMLR, 2020, pp. 1597–1607.

[12] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, 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.

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

[14] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, Armand Joulin, Unsupervised learning of visual features by contrasting cluster assignments, Adv. Neural Inf. Process. Syst. 33 (2020) 9912–9924.

[15] Alexander Ke, William Ellsworth, Oishi Banerjee, Andrew Y Ng, Pranav Rajpurkar, CheXtransfer: Performance and parameter efficiency of ImageNet models for chest X-Ray interpretation, in: Proceedings of the Conference on Health, Inference, and Learning, 2021, pp. 116–124.

[16] Almas Begum, V Dhilip Kumar, Junaid Asghar, D Hemalatha, G Arulkumaran, A combined deep CNN: LSTM with a random forest approach for breast cancer diagnosis, Complexity 2022 (2022).

[17] Ari M. Ali, Aree A. Mohammed, Improving classification accuracy for prostate cancer using noise removal filter and deep learning technique, Multimedia Tools Appl. 81 (6) (2022) 8653–8669.

[18] Li-Yin Ye, Xiao-Yan Miao, Wan-Song Cai, Wan-Jiang Xu, Medical image diagnosis of prostate tumor based on PSP-net+ VGG16 deep learning network, Comput. Methods Programs Biomed. 221 (2022) 106770.

[19] Lourdes Duran-Lopez, Juan P Dominguez-Morales, Daniel Gutierrez-Galan, Antonio Rios-Navarro, Angel Jimenez-Fernandez, Saturnino Vicente-Diaz, Alejandro Linares-Barranco, Wide & deep neural network model for patch aggregation in CNN-based prostate cancer detection systems, Comput. Biol. Med. 136 (2021) 104743.

[20] Xu Lu, Shaohui Wu, Zhiwei Xiao, Xiongwei Huang, An enhanced multiscale generation and depth-perceptual loss-based super-resolution network for prostate ultrasound images, Meas. Sci. Technol. 34 (2) (2022) 024002.

[21] Mustafa Koc, Suat Kamil Sut, Ihsan Serhatlioglu, Mehmet Baygin, Turker Tuncer, Automatic prostate cancer detection model based on ensemble vggnet feature generation and NCA feature selection using magnetic resonance images, Multimedia Tools Appl. 81 (5) (2022) 7125–7144.

[22] Haibo Yang, GuangYu Wu, Dinggang Shen, Shu Liao, Automatic prostate cancer detection on multi-parametric mri with hierarchical weakly supervised learning, in: 2021 IEEE 18th International Symposium on Biomedical Imaging, ISBI, IEEE, 2021, pp. 316–319.

[23] Longlong Jing, Yingli Tian, Self-supervised visual feature learning with deep neural networks: A survey, IEEE Trans. Pattern Anal. Mach. Intell. 43 (11) (2020) 4037–4058.

[24] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, Dahua Lin, Unsupervised feature learning via non-parametric instance discrimination, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3733–3742.

[25] Mang Ye, Xu Zhang, Pong C. Yuen, Shih-Fu Chang, Unsupervised embedding learning via invariant and spreading instance feature, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 6210–6219.

[26] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al., Bootstrap your own latent-a new approach to self-supervised learning, Adv. Neural Inf. Process. Syst. 33 (2020) 21271–21284.

[27] Yutong Xie, Jianpeng Zhang, Zehui Liao, Yong Xia, Chunhua Shen, PGL: Prior-guided local self-supervised learning for 3D medical image segmentation, 2020, arXiv preprint arXiv:2011.12640.

[28] Lang Huang, Shan You, Mingkai Zheng, Fei Wang, Chen Qian, Toshihiko Yamasaki, Learning where to learn in cross-view self-supervised learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 14451–14460.

[29] Xiangyu Peng, Kai Wang, Zheng Zhu, Mang Wang, Yang You, Crafting better contrastive views for siamese representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 16031–16040.

[30] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 618–626.

[31] Agisilaos Chartsias, Shan Gao, Angela Mumith, Jorge Oliveira, Kanwal Bhatia, Bernhard Kainz, Arian Beqiri, Contrastive learning for view classification of echocardiograms, in: Simplifying Medical Ultrasound: Second International Workshop, ASMUS 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings 2, Springer, 2021, pp. 149–158.

[32] Xiaohan Xing, Yuenan Hou, Hang Li, Yixuan Yuan, Hongsheng Li, Max Q-H Meng, Categorical relation-preserving contrastive knowledge distillation for medical image classification, in: Medical Image Computing and Computer Assisted Intervention, MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24, Springer, 2021, pp. 163–173.

[33] Yijin Huang, Li Lin, Pujin Cheng, Junyan Lyu, Xiaoying Tang, Lesion-based contrastive learning for diabetic retinopathy grading from fundus images, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2021, pp. 113–123.

[34] Prasanna Porwal, Samiksha Pachade, Ravi Kamble, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabuddhe, Fabrice Meriaudeau, Indian diabetic retinopathy image dataset (idrid): A database for diabetic retinopathy screening research, Data 3 (3) (2018) 25.

[35] Ben Graham, Kaggle Diabetic Retinopathy Detection Competition Report, University of Warwick, 2015, pp. 24–26.

[36] Hong-Yu Zhou, Chixiang Lu, Sibei Yang, Xiaoguang Han, Yizhou Yu, Preservational learning improves self-supervised medical image models by reconstructing diverse contexts, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 3499–3509.

[37] Nanqing Dong, Irina Voiculescu, Federated contrastive learning for decentralized unlabeled medical images, in: Medical Image Computing and Computer Assisted Intervention, MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24, Springer, 2021, pp. 378–387.

[38] Alvaro Fernandez-Quilez, Trygve Eftestøl, Svein Reidar Kjosavik, Morten Goodwin, Ketil Oppedal, Contrasting axial T2W MRI for prostate cancer triage: A self-supervised learning approach, in: 2022 IEEE 19th International Symposium on Biomedical Imaging, ISBI, IEEE, 2022, pp. 1–5.

[39] Jun Fu, Weisheng Li, Jiao Du, Yuping Huang, A multiscale residual pyramid attention network for medical image fusion, Biomed. Signal Process. Control 66 (2021) 102488.

[40] Rui Xu, Zhen Cong, Xinchen Ye, Yasushi Hirano, Shoji Kido, Tomoko Gyobu, Yutaka Kawata, Osamu Honda, Noriyuki Tomiyama, Pulmonary textures classification via a multi-scale attention network, IEEE J. Biomed. Health Inform. 24 (7) (2019) 2041–2052.

[41] Yimian Dai, Fabian Gieseke, Stefan Oehmcke, Yiquan Wu, Kobus Barnard, Attentional feature fusion, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 3560–3569.

[42] Ashish Sinha, Jose Dolz, Multi-scale self-guided attention for medical image segmentation, IEEE J. Biomed. Health Inform. 25 (1) (2020) 121–130.

[43] Yuchen Yuan, Lei Zhang, Lituan Wang, Haiying Huang, Multi-level attention network for retinal vessel segmentation, IEEE J. Biomed. Health Inf. 26 (1) (2021) 312–323.

[44] Olaf Ronneberger, Philipp Fischer, Thomas Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015, pp. 234–241.

[45] Jiawei Sun, Chunying Li, Zhengda Lu, Mu He, Tong Zhao, Xiaoqin Li, Liugang Gao, Kai Xie, Tao Lin, Jianfeng Sui, et al., Tnsnet: Thyroid nodule segmentation in ultrasound imaging using soft shape supervision, Comput. Methods Programs Biomed. 215 (2022) 106600.

[46] Yanlin Wu, Guanglei Wang, Zhongyang Wang, Hongrui Wang, Yan Li, DI-unet: Dimensional interaction self-attention for medical image segmentation, Biomed. Signal Process. Control 78 (2022) 103896.

[47] Sanguk Park, Minyoung Chung, Cardiac segmentation on ct images through shape-aware contour attentions, Comput. Biol. Med. 147 (2022) 105782.

[48] Yibo Feng, Xu Yang, Dawei Qiu, Huan Zhang, Dejian Wei, Jing Liu, Pcxrnet: Pneumonia diagnosis from chest X-Ray images using condense attention block and multiconvolution attention block, IEEE J. Biomed. Health Inf. 26 (4) (2022) 1484–1495.

[49] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, Hanqing Lu, Dual attention network for scene segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3146–3154.

[50] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[51] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, Han Hu, Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 16684–16693.

[52] Zhiyong Liu, Chuan Yang, Jun Huang, Shaopeng Liu, Yumin Zhuo, Xu Lu, Deep learning framework based on integration of S-mask R-CNN and inception-v3 for ultrasound image-aided diagnosis of prostate cancer, Future Gener. Comput. Syst. 114 (2021) 358–367.

[53] Ilya Loshchilov, Frank Hutter, Sgdr: Stochastic gradient descent with warm restarts, 2016, arXiv preprint arXiv:1608.03983.