HiCo: Hierarchical Contrastive Learning for Ultrasound Video Model Pretraining

Chunhui Zhang[1,2][0000−0002−9017−1828], Yixiong Chen[3,4][0000−0003−0268−076X], Li Liu[3,4,⋆][0000−0002−4497−0135], Qiong Liu[2][0000−0002−5808−2761], and Xi Zhou[2,1][0000−0001−9943−5482]

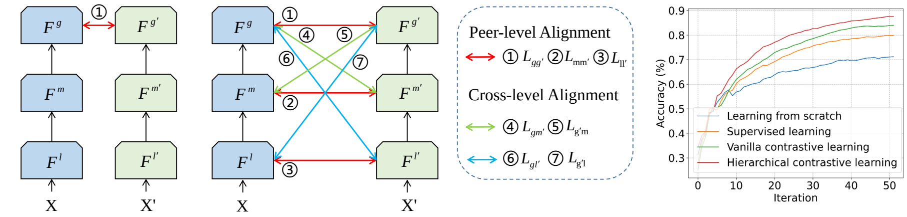
[1] Shanghai Jiaotong University, 200240 Shanghai, China
[2] CloudWalk Technology Co., Ltd, 201203 Shanghai, China
[3] The Chinese University of Hong Kong (Shenzhen), 518172 Shenzhen, China
`liuli@cuhk.edu.cn`
[4] Shenzhen Research Institute of Big Data, 518172 Shenzhen, China

Abstract. The self-supervised ultrasound (US) video model pretraining can use a small amount of labeled data to achieve one of the most promising results on US diagnosis. However, it does not take full advantage of multi-level knowledge for learning deep neural networks (DNNs), and thus is difficult to learn transferable feature representations. This work proposes a hierarchical contrastive learning (HiCo) method to improve the transferability for the US video model pretraining. HiCo introduces both peer-level semantic alignment and cross-level semantic alignment to facilitate the interaction between different semantic levels, which can effectively accelerate the convergence speed, leading to better generalization and adaptation of the learned model. Additionally, a softened objective function is implemented by smoothing the hard labels, which can alleviate the negative effect caused by local similarities of images between different classes. Experiments with HiCo on five datasets demonstrate its favorable results over state-of-the-art approaches. The source code of this work is publicly available at https://github.com/983632847/HiCo.

1 Introduction

Thanks to the cost-effectiveness, safety, and portability, combined with a reasonable sensitivity to a wide variety of pathologies, ultrasound (US) has become one of the most common medical imaging techniques in clinical diagnosis [1]. To mitigate sonographers' reading burden and improve diagnosis efficiency, automatic US analysis using deep learning is becoming popular [2–5]. In the past decades, a successful practice is to train a deep neural network (DNN) on a large number of well-labeled US images within the supervised learning paradigm [1, 6]. However, annotations of US images and videos can be expensive to obtain and sometimes infeasible to access because of the expertise requirements and time-consuming reading, which motivates the development of US diagnosis that requires few or even no manual annotations.

[⋆] Corresponding author.



(a) Vanilla contrastive learning (b) Hierarchical contrastive learning (c) Convergence speed

Fig. 1. Motivation of hierarchical contrastive learning. Unlike (a) vanilla contrastive learning, our (b) hierarchical contrastive learning can fully take advantage of both peer-level and cross-level information. Thus, (c) the pretraining model from our proposed hierarchical contrastive learning can accelerate the convergence speed, which is much better than learning from scratch, supervised learning, and vanilla contrastive learning.

In recent years, pretraining combined with fine-tuning has attracted great attention because it can transfer knowledge learned on large amounts of unlabeled or weakly labeled data to downstream tasks, especially when the amount of labeled data is limited. This has also profoundly affected the field of US diagnosis, which started to pretrain models from massive unlabeled US data according to a pretext task. To learn meaningful and strong representations, the US video pretraining methods are designed to correct the order of a reshuffled video clip, predict the geometric transformation applied to the video clip or colorize a grayscale image to its color version equivalent [7, 8]. Inspired by the powerful ability of contrastive learning (CL) [9, 10] in computer vision, some recent studies propose to learn US video representations by CL [11, 3], and showed a powerful learning capability [12, 11]. However, most of the existing US video pretraining methods following the vanilla contrastive learning setting [10, 13], only use the output of a certain layer of a DNN for contrast (see Fig. 1(a)). Although the CL methods are usually better than learning from scratch and supervised learning, the lack of multi-level information interaction will inevitably degrade the transferability of pretrained models [3, 14].

To address the above issue, we first propose a hierarchical contrastive learning (HiCo) method for US video model pretraining. The main motivation is to design a *feature-based* peer-level and cross-level semantic alignment method (see Fig. 1(b)) to improve the efficiency of learning and enhance the ability of feature representation. Specially, based on the assumption that the top layer of a DNN has strong semantic information, and the bottom layer has high-resolution local information (*e.g.*, texture and shape) [15], we design a joint learning task to force the model to learn multi-level semantic representations during the CL process: minimize the peer-level semantic alignment loss (*i.e.*, ① global CL loss, ② medium CL loss, and ③ local CL loss) and cross-level semantic alignment loss (*i.e.*, ④, ⑤ global-medium CL losses, and ⑥, ⑦ global-local CL losses) simultaneously. Intuitively, our framework can greatly improve the convergence speed of the model (*i.e.*, providing a good initialized model for downstream tasks) (see

Fig. 1(c)), due to the sufficient interaction of peer-level and cross-level information. Different from existing methods [16–20], this work assumes that the knowledge inside the backbone is sufficient but underutilized, so that simple yet effective peer-level and cross-level semantic alignments can be used to enhance feature representation other than designing a complex structure. In addition, medical images from different classes/lesions may have significant local similarities (*e.g.*, normal and infected individuals have similar regions of tissues and organs unrelated to disease), which is more severe than natural images. Thus, we follow the popular label smoothing strategy to design a *batch-based* softened objective function during the pretraining to avoid the model being over-confident, which alleviates the negative effect caused by local similarities.

The main contributions of this work can be summarized as follows:

1) We propose a novel hierarchical contrastive learning method for US video model pretraining, which can make full use of the multi-level knowledge inside a DNN via peer-level semantic alignment and cross-level semantic alignment.

2) We soften one-hot labels during the pretraining process to avoid the model being over-confident, alleviating the negative effect caused by local similarities of images between different classes.

3) Experiments on five downstream tasks demonstrate the effectiveness of our approach in learning transferable representations.

2 Related Work

We first review related works on supervised learning for US diagnosis and then discuss the self-supervised representation learning.

2.1 US Diagnosis

With the rise of deep learning in computer vision, supervised learning became the most common strategy in US diagnosis with DNN [1, 3, 21–23]. In the last decades, numerous datasets and methods have been introduced for US image classification [24], detection [25] and segmentation [26] tasks. For example, some US image datasets with labeled data were designed for breast cancer classification [27, 28], breast US lesions detection [25], diagnosis of malignant thyroid nodule [29, 30], and automated measurement of the fetal head circumference [31]. At the same time, many deep learning approaches have been done on lung US [32, 33], B-line detection or quantification [34, 35], pleural line extraction [36], and subpleural pulmonary lesions [37]. Compared with image-based datasets, recent video-based US datasets [1, 3] are becoming much richer and can provide more diverse categories and data modalities (*e.g.*, convex and linear probe US images [3]). Thus, many works are focused on video-based US diagnosis within the supervised learning paradigm. In [1], a frame-based model was proposed to correctly distinguish COVID-19 lung US videos from healthy and bacterial pneumonia data. Other works focus on quality assessment for medical US video compressing [38], localizing target structures [39], or describing US video content [2]. Until recently, many advanced DNNs (*e.g.*, UNet [40], DeepLab [41, 42], Transformer [43]), and technologies (*e.g.*, neural architecture search [44], reinforcement learning [45], meta-learning [46]) have brought great advances in supervised learning for US diagnosis. Unfortunately, US diagnosis using supervised learning highly relies on large-scale labeled, often expensive medical datasets.

2.2 Self-supervised Learning

Recently, many self-supervised learning methods for visual feature representation learning have been developed without using any human-annotated labels [47–49]. Existing self-supervised learning methods can be divided into two main categories, *i.e.*, learning via pretext tasks and CL. A wide range of pretext tasks have been proposed to facilitate the development of self-supervised learning. Examples include solving jigsaw puzzles [50], colorization [8], image context restoration [51], and relative patch prediction [52]. However, many of these tasks rely on ad-hoc heuristics that could limit the generalization and robustness of learned feature representations for downstream tasks [13, 10]. The CL has emerged as the front-runner for self-supervision representation learning and has demonstrated remarkable performance on downstream tasks. Unlike learning via pretext tasks, CL is a discriminative approach that aims at grouping similar positive samples closer and repelling negative samples. To achieve this, a similarity metric is used to measure how close two feature embeddings are. For computer vision tasks, a standard loss function, *i.e.*, Noise-Contrastive Estimation loss (InfoNCE) [53], is evaluated based on the feature representations of images extracted from a backbone network (*e.g.*, ResNet [22]). Most successful CL approaches are focused on studying effective contrastive loss, generation of positive and negative pairs, and sampling method [10, 9]. SimCLR [10] is a simple framework for CL of visual representations with strong data augmentations and a large training batch size. MoCo [9] builds a dynamic dictionary with a queue and a moving-averaged encoder. Other works explores learning without negative samples [54, 55], and incorporating self-supervised learning with visual transformers [56], *etc.*

Considering the superior performance of contrastive self-supervised learning in computer vision and medical imaging tasks, this work follows the line of CL. First, we propose both peer-level and cross-level alignments to speed up the convergence of the model learning, compared with the existing CL methods, which usually use the output of a certain layer of the network for contrast (see Fig. 1). Second, we design a softened objective function to facilitate the CL by addressing the negative effect of local similarities between different classes.

3 Hierarchical Contrastive Learning

In this section, we present our HiCo approach for US video model pretraining. To this end, we first introduce the preliminary of CL, after that present the peer-level semantic alignment and cross-level semantic alignment, and then describe the softened objective function. The framework of HiCo is illustrated in Fig. 2.

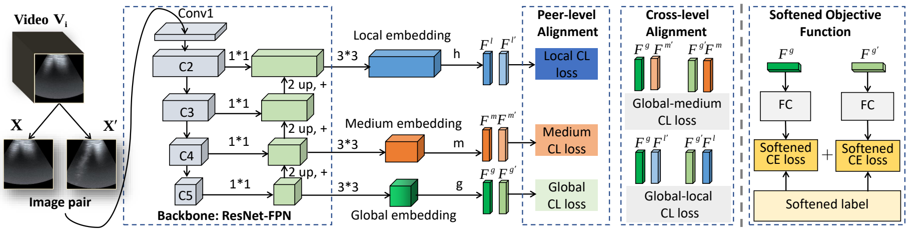


Fig. 2. Overall framework of the proposed HiCo, which consists of peer-level semantic alignment, cross-level semantic alignment, and softened objective function. 1) We extract two images from each US video as a positive sample pair. 2) We use ResNet-FPN as the backbone to obtain the local, medium and global embeddings, followed by three projection heads h, m and g. 3) The entire network is optimized by minimizing peer-level semantic alignment loss (*i.e.*, local CL loss, medium CL loss and global CL loss), cross-level semantic alignment loss (*i.e.*, global-medium CL loss and global-local CL loss), and the softened CE loss.

3.1 Preliminary

The vanilla contrastive learning learns a global feature encoder Φ and a projection head g that map the image X into a feature vector $\mathbf{F} = g(\Phi(X)) \in \mathbb{R}_d$ by minimizing the InfoNCE loss [53]:

$$\mathscr{L}_{nce}(\mathbf{F}_i, \mathbf{F}_j) = -\log \frac{exp(sim(\mathbf{F}_i, \mathbf{F}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} exp(sim(\mathbf{F}_i, \mathbf{F}_k)/\tau)}. \tag{1}$$

where $(\mathbf{F}_i, \mathbf{F}_j)$ are the global feature vectors of the two views of image X, N is the batch size. $sim(\mathbf{F}_i, \mathbf{F}_j) = \mathbf{F}_i \cdot \mathbf{F}_j/(||\mathbf{F}_i||||\mathbf{F}_j||)$ denotes the cosine similarity, $\mathbb{1}_{k \neq i} \in \{0, 1\}$ is an indicator function evaluating to 1 iff $k \neq i$, and τ is a tuning temperature parameter.

In practice, Φ is a DNN (*e.g.*, ResNet [22]) and the objective function is minimized using stochastic gradient descent. The two feature vectors of the one image are a positive pair, while the other $2(N-1)$ examples within a mini-batch are treated as negative examples. The positive and negative contrastive feature pairs are usually the global representations (*e.g.*, the output from the last layer of a DNN) that lack multi-scale information from different feature layers, which are important for downstream tasks. In addition, having a large number of negative samples is critical while minimizing the InfoNCE loss. Hence, a large batch size (*e.g.*, 4096 in SimCLR [10]) is required during the pretraining.

To address the above problem, we explore multi-level semantic alignments for CL, peer-level and cross-level ones (see Fig.2). Specifically, we use ResNet-FPN as the backbone. For each image pair, we extract two local feature vectors ($\mathbf{F}^l$ and $\mathbf{F}^{l'}$), two medium feature vectors ($\mathbf{F}^m$ and $\mathbf{F}^{m'}$) and two global feature vectors ($\mathbf{F}^g$ and $\mathbf{F}^{g'}$) from Conv2 (C2), Conv4 (C4) and Conv5 (C5), respectively. We then optimize the peer-level and cross-level alignments simultaneously.

3.2 Peer-level Semantic Alignment

Fine-grained Contrast. The $\mathbf{F}^l$ and $\mathbf{F}^{l'}$ encode the fine-grained local information (*e.g.*, edges and shapes) of original images. Such fine-grained information is useful for US diagnosis, but is usually ignored in existing CL algorithms. To leverage the fine-grained information, we define the local CL loss $\mathcal{L}_{ll'}^{local}$ as

$$\mathcal{L}_{ll'}^{local} = \frac{1}{2N}\sum_{i=1}^{N}(\mathscr{L}_{nce}(\mathbf{F}_i^l, \mathbf{F}_i^{l'}) + \mathscr{L}_{nce}(\mathbf{F}_i^{l'}, \mathbf{F}_i^l)), \tag{2}$$

where N denotes the batch size, $\mathscr{L}_{nce}(\cdot)$ is the InfoNCE loss [53].

Medium-grained Contrast. Considering that $\mathbf{F}^m$ and $\mathbf{F}^{m'}$ capture medium-grained information of original images, we therefore define the medium CL loss $\mathcal{L}_{mm'}^{medium}$ as

$$\mathcal{L}_{mm'}^{medium} = \frac{1}{2N}\sum_{i=1}^{N}(\mathscr{L}_{nce}(\mathbf{F}_i^m, \mathbf{F}_i^{m'}) + \mathscr{L}_{nce}(\mathbf{F}_i^{m'}, \mathbf{F}_i^m)), \tag{3}$$

Notably, we find that medium-grained information demonstrate complementary superiority relative to fine-grained and global information, further improving model performance (see Section 4.2 Table 1).

Coarse-grained Contrast. Owing to $\mathbf{F}^g$ and $\mathbf{F}^{g'}$ capture coarse-grained global information of original images, we hope to reach a consensus among their representations by maximizing the similarity between global embeddings from the same video, while minimizing the similarity between the global embeddings from different videos. Thus, the global CL loss $\mathcal{L}_{gg'}^{global}$ can be defined as

$$\mathcal{L}_{gg'}^{global} = \frac{1}{2N}\sum_{i=1}^{N}(\mathscr{L}_{nce}(\mathbf{F}_i^g, \mathbf{F}_i^{g'}) + \mathscr{L}_{nce}(\mathbf{F}_i^{g'}, \mathbf{F}_i^g)), \tag{4}$$

3.3 Cross-level Semantic Alignment

Global-local Contrast. We regard the global feature vector $\mathbf{F}^g$ as the *anchor* of the local feature vector $\mathbf{F}^{l'}$, because it contains the global semantic information of the original image and shares some semantic content with the local feature vector. Thus, we define the global-local objective $\mathcal{L}_{gl'}$ to make the local feature vectors move closer to the global ones as

$$\mathcal{L}_{gl'} = \frac{1}{2N}\sum_{i=1}^{N}(\mathscr{L}_{nce}(\mathbf{F}_i^g, \mathbf{F}_i^{l'}) + \mathscr{L}_{nce}(\mathbf{F}_i^{l'}, \mathbf{F}_i^g)), \tag{5}$$

The $\mathcal{L}_{g'l}$ can be calculated similarly. Then, the global-local CL loss can be written as $\mathcal{L}_{local}^{global} = \mathcal{L}_{gl'} + \mathcal{L}_{g'l}$.

Algorithm 1 Hierarchical Contrastive Learning

```
Input: US videos $\mathbf{V}$, backbone $\Phi$, projection heads $g, m, h$, linear classifier $f_\theta$, hyper-parameters $\lambda, \alpha, \beta$, max epoch $e_{max}$, batch size $N$.
Output: pretrained backbone $\Phi$.
random initialize $\Phi$, $g, m, h$, and $f_\theta$
for $e = 1, 2, ..., e_{max}$ do
    for random sampled US videos $\{\mathbf{V}_i\}_{i=1}^N$ do
        # extract two images from each US video as a positive sample pair.
        random sample $\{(\mathbf{x}_i^{(1)}, \mathbf{x}_i^{(2)}), \mathbf{y}_i\}_{i=1}^N$ into mini-batch
        for $i \in \{1, ..., N\}$ do
            # Augment image pair for each US video.
            random cropping, resizing, flipping, and color jitter
            # Get local, medium and global embeddings.
            $\mathbf{F}_i^l = h(\Phi(\mathbf{x}_i^{(1)}))$, $\mathbf{F}_i^{l'} = h(\Phi(\mathbf{x}_i^{(2)}))$;
            $\mathbf{F}_i^m = m(\Phi(\mathbf{x}_i^{(1)}))$, $\mathbf{F}_i^{m'} = m(\Phi(\mathbf{x}_i^{(2)}))$;
            $\mathbf{F}_i^g = g(\Phi(\mathbf{x}_i^{(1)}))$, $\mathbf{F}_i^{g'} = g(\Phi(\mathbf{x}_i^{(2)}))$;
            # Get outputs of the linear classifier.
            $\mathbf{o}_i = f_\theta(\mathbf{F}_i^g)$, $\mathbf{o}_i' = f_\theta(\mathbf{F}_i^{g'})$
    # peer-level semantic alignment
    for $i \in \{1, ..., N\}$ do
        calculate the local CL loss $\mathcal{L}_{ll'}^{local}$ by Eq. (2)
        calculate the medium CL loss $\mathcal{L}_{mm'}^{medium}$ by Eq. (3)
        calculate the global CL loss $\mathcal{L}_{gg'}^{global}$ by Eq. (4)
    # cross-level semantic alignment
    for $i \in \{1, ..., N\}$ do
        calculate the global-local CL loss $\mathcal{L}_{local}^{global}$ by Eq. (5)
        calculate the global-medium CL loss $\mathcal{L}_{medium}^{global}$
    calculate the overall CL loss $\mathcal{L}^{con}$ by Eq. (6)
    calculate the softened CE loss $\mathcal{L}^{soften}$ by Eq. (8)
    $\mathcal{L} = \mathcal{L}^{con} + \beta\mathcal{L}^{soften}$
    update $\Phi$, $g, m, h$, and $f_\theta$ through gradient descent
return pretrained $\Phi$, discard $g, m, h$, and $f_\theta$
```

Global-medium Contrast. Similar to the global-local CL loss, the global-medium CL loss can be written as $\mathcal{L}_{medium}^{global} = \mathcal{L}_{gm'} + \mathcal{L}_{g'm}$.

Therefore, the overall CL loss of HiCo is formulated as

$$\mathcal{L}^{con} = \lambda(\mathcal{L}_{ll'}^{local} + \mathcal{L}_{mm'}^{medium} + \mathcal{L}_{gg'}^{global}) + (1 - \lambda)(\mathcal{L}_{local}^{global} + \mathcal{L}_{medium}^{global}), \quad (6)$$

where λ is a trade-off coefficient, the first three terms $\mathcal{L}_{ll'}^{local}$, $\mathcal{L}_{mm'}^{medium}$ and $\mathcal{L}_{gg'}^{global}$ represent the peer-level semantic alignment objective functions, while the last two terms $\mathcal{L}_{local}^{global}$ and $\mathcal{L}_{medium}^{global}$ represent the cross-level semantic alignment objective functions.

3.4 Softened Objective Function

The one-hot label assumes there is absolutely no similarity between different classes. However, in medical imaging, the images from different classes may have some local similarities (*e.g.*, the tissues and organs unrelated to diseases). Thus, we propose the softened objective function to alleviate the negative effect caused by local similarities. We first define the corresponding softened label $\widetilde{\mathbf{y}}_i$ as

$$\widetilde{\mathbf{y}}_i = (1-\alpha)\mathbf{y}_i + \alpha/(N-1), \tag{7}$$

where $\mathbf{y}_i$ is the original one-hot label, α is a smoothing hyper-parameter, and N is the number of videos in a training batch. Then the softened cross-entropy (CE) loss $\mathcal{L}^{soften}$ with corresponding softened label $\widetilde{\mathbf{y}}_i$ is formulated as

$$\mathcal{L}^{soften} = \frac{1}{2N}\sum_{i=1}^{N}(CE(\mathbf{o}_i, \widetilde{\mathbf{y}}_i) + CE(\mathbf{o}'_i, \widetilde{\mathbf{y}}_i)), \tag{8}$$

where $\mathbf{o}_i = f_\theta(\mathbf{F}_i^g)$, $\mathbf{o}'_i = f_\theta(\mathbf{F}_i^{g'})$, and f_θ is a linear classifier.

Finally, the total loss can be written as

$$\mathcal{L} = \mathcal{L}^{con} + \beta\mathcal{L}^{soften}, \tag{9}$$

where the parameter β is used to balance the total CL loss and softened CE loss. The whole algorithm is summarized in Algorithm 1.

4 Experiments

4.1 Experimental Settings

Network Architective. In our experiments, we apply the widely used ResNet18-FPN [57] network as the backbone. Conv2 to Conv5 are followed by a convolution layer with kernel size of 1*1 to obtain intermediate feature maps. In the FPN structure, we double upsampling the intermediate feature maps, and then add them to the intermediate feature maps of the previous layer. The intermediate feature maps are followed by a convolution layer with kernel size of 3*3 to obtain local embedding, medium embedding, and global embedding, respectively. All convolution layers are followed by batch normalization and ReLU. The projection heads h, m, and g are all 1-layer MLP. After the projection heads, the local, medium, and global embeddings are reduced to 256-dimensional feature vectors for CL tasks. The linear classifier is a fully connected (FC) layer.

Pretraining Details. We use the US-4 [3] video dataset (lung and liver) for pretraining and fine-tune the last 3 layers of pretrained models on various downstream tasks to evaluate the transferability of the proposed US video pretraining models. During the pretraining process, the input images are randomly cropped and resized to 224*224, followed by random flipping and color jitter. The pretraining epoch and batch size are set to 300 and 32, respectively. The parameters of models are obtained by optimizing the loss functions via an Adam optimizer with a learning rate 3×10^{-4} and a weight decay rate 10^{-4}. Following the popular CL evaluations [10], the backbone is used for fine-tuning on downstream tasks, projection heads (h, m, and g) and linear classifier (f_θ) are discarded when the pretraining is completed. The $\tau = 0.5$ is a tuning temperature parameter as in SimCLR [10]. We empirically set $\lambda = 0.5$, indicating that peer-level semantic alignment and cross-level semantic alignment have equal weights in the CL loss. The smoothing parameter α set to 0.2 indicates slight label smoothing. The β is empirically set to 0.2, indicating that the CL loss dominates the total loss. All experiments were implemented using PyTorch and a single RTX 3090 GPU.

Downstream Datasets. We fine-tune our pretrained backbones on four US datasets (POCUS [1], Thyroid US [29], and BUSI-BUI [27, 28] joint dataset), and a chest X-ray dataset (COVID-Xray-5k [58]) to evaluate the transferability of our pretraining models. For fair comparisons, all fine-tuning results on downstream datasets are obtained with 5-fold cross-validation. The POCUS is a lung convex probe US dataset for pneumonia detection that contains 2116 frames across 140 videos from three categories (COVID-19, bacterial pneumonia and the regular). The BUSI contains 780 breast tumor US images from three classes (the normal, benign and malignant), while BUI consists of 250 breast cancer US images with 100 benign and 150 malignant. Thyroid US [29] dataset contains thyroid images with 61 benign and 288 malignant. Note that the BUSI, BUI and Thyroid US datasets are collected with linear probes. The COVID-Xray-5k is a chest X-ray dataset that contains 2084 training and 3100 test images from two classes (COVID-19 and the normal). In our fine-tuning experiments, the learning rate, weight decay rate and epoch are set to 10^{-2}, 10^{-4} and 30, respectively. The performance is assessed with Precision, Recall, Accuracy or F1 score.

4.2 Ablation Studies

In this section, we verify the effectiveness of each component in our approach on the downstream POCUS pneumonia detection task, and all the models are pretrained on the US-4 dataset.

Peer-level Semantic Alignment. The impact of peer-level semantic alignment is summarized in Table 1. We reimplement the self-supervised method SimCLR [10] with ResNet18 (w/o FPN) as our baseline. We can find that the baseline cannot achieve satisfying performance (85.6%, 87.0% and 86.9% in terms of Precision, Recall and Accuracy, respectively). The effectiveness of fine-grained contrast, medium-grained contrast and coarse-grained contrast can be verified by comparing backbones pretrained using $\mathcal{L}_{ll'}^{local}$, $\mathcal{L}_{mm'}^{medium}$, $\mathcal{L}_{gg'}^{global}$ with the baseline, which contribute to the absolute performance gains of 1.4%, 2.7% and 3.2% in terms of Accuracy. In addition, better or comparable results can be achieved

Table 1. Impact of peer-level semantic alignment. The models are pretrained on US-4 and fine-tuned on POCUS dataset.

$\mathcal{L}_{ll'}^{local}$	$\mathcal{L}_{mm'}^{medium}$	$\mathcal{L}_{gg'}^{global}$	Precision (%)	Recall (%)	Accuracy (%)
			85.6	87.0	86.9
✓			87.6	87.5	88.3
	✓		87.9	90.2	89.6
		✓	90.5	89.8	90.1
✓	✓		89.3	89.5	90.1
✓		✓	91.0	90.7	90.5
	✓	✓	91.9	92.3	91.9
✓	✓	✓	91.3	92.3	92.0

Table 2. Impact of cross-level semantic alignment. The models are pretrained on US-4 and fine-tuned on the POCUS dataset.

$\mathcal{L}_{medium}^{global}$	$\mathcal{L}_{local}^{global}$	Precision (%)	Recall (%)	Accuracy (%)
		85.6	87.0	86.9
✓		88.0	89.6	89.4
	✓	90.4	91.1	90.8
✓	✓	89.4	90.8	90.9

when using two contrasts (*e.g.*, coarse-grained contrast and medium-grained contrast) than a single contrast (*e.g.*, coarse-grained contrast). Considering the excellent performance of backbones when using only the global CL loss (*i.e.*, an Accuracy of 90.1%), and using both the global CL loss and the medium CL loss (*i.e.*, an Accuracy of 91.9%), we argue that coarse-grained global information is very important to improve the transferability of pretrained US models. The best performance is achieved when all three peer-level semantic contrasts are used.

Cross-level Semantic Alignment. The impact of cross-level semantic alignment is summarized in Table 2. In previous experiments, we find that global information is pivotal to CL tasks. Therefore, when performing cross-level semantic alignment, we regard the global feature as an *anchor*, and only consider aligning the local features and the middle-level features with the global features (*i.e.*, global-local contrast and global-medium contrast). From Table 2, we can observe that consistent performance gains are achieved by conducting global-local contrast and global-medium contrast in terms of Precision, Recall and Accuracy. When global-local contrast and global-medium contrast are performed at the same time, our pretrained model can achieve the best performance in terms of Accuracy (*i.e.*, 90.9%). However, using the proposed peer-level semantic alignment can achieve the best Accuracy of 92.0% as shown in Table 1. Therefore, it is difficult to learn better transferable representations by using cross-level semantic alignment alone. Next, we will demonstrate that using the peer-level semantic alignment and cross-level semantic alignment at the same time can bring significant performance gains.

Table 3. Ablation study of each component in our approach. The models are pretrained on US-4 and fine-tuned on POCUS dataset.

$\mathcal{L}^{local}_{ll'}$	$\mathcal{L}^{medium}_{mm'}$	$\mathcal{L}^{global}_{gg'}$	$\mathcal{L}^{global}_{medium}$	$\mathcal{L}^{global}_{local}$	$\mathcal{L}^{soften}$	Precision (%)	Recall (%)	Accuracy (%)
						85.6	87.0	86.9
✓						87.6	87.5	88.3
✓	✓					89.3	89.5	90.1
✓	✓	✓				91.3	92.3	92.0
✓	✓	✓	✓			92.2	93.2	93.2
✓	✓	✓	✓	✓		93.5	94.2	94.4
✓	✓	✓	✓	✓	✓	94.6	94.9	94.7

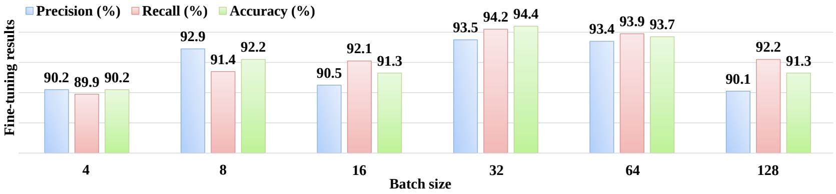


Fig. 3. Impact of batch size. Fine-tuning results obtained on POCUS dataset with different batch sizes.

Ablation Study of Each Component. The impact of each component in our approach is summarized in Table 3. We can see that each component can bring a certain performance improvement to our final method, which verifies the effectiveness of each component. An interesting observation is that our method (an Accuracy of 94.4%) without using labels can surpass the current state-of-the-art semi-supervised method USCL (an Accuracy of 94.2%) and supervised method (an Accuracy of 85.0%) on POCUS dataset (see Tables 3 and 4). This result demonstrates the superiority of peer-level semantic alignment and cross-level semantic alignment for learning transferable representations. In addition, the softened objective function can further facilitate the performance of our approach (from 94.4% to 94.7% in terms of Accuracy).

Impact of Batch Size. The effect of batch size is shown in Fig. 3. We can observe that as the batch size increases, the overall performance demonstrates a trend from rising to decline, where the best overall performance is achieved when the batch size is 32. Compared with existing state-of-the-art CL algorithms that rely on large batch sizes (more negative samples), *e.g.*, 1024 in MoCo [9], 2048 in SimCLR [10], to achieve convergence, our proposed method is easier to optimize. In this way, we can train the model in fewer epochs and steps to obtain a given accuracy. We argue that the multi-level contrast promotes the effective and sufficient interaction of information from different layers of a DNN, thus we can use a smaller batch size.

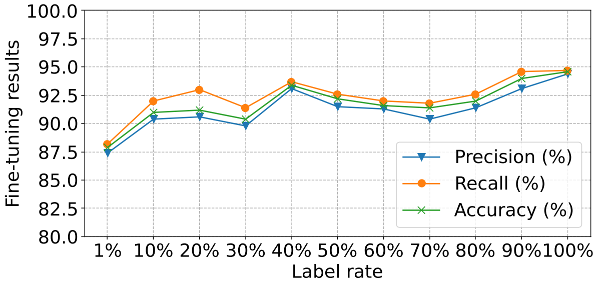


Fig. 4. Impact of the label rate. Fine-tuning results obtained on POCUS dataset with different label rates.

Impact of Label Rate. The effect of label rate is shown in Fig. 4. We pretrain eleven models on US-4 dataset with different label rates. We find that the performance of models gradually improve with the increase of label, as more label can bring stronger supervision signals to promote the process of pretraining.

4.3 Comparison with State-of-the-art Methods

To verify the effectiveness of our approach, we compare the proposed HiCo with supervised ResNet18 backbones (*i.e.*, "ImageNet" pretrained on ImageNet dataset, "Supervised" pretrained on US-4 dataset), and other backbones pretrained on US-4 dataset with semi-supervised methods (*i.e.*, Temporal Ensembling (TE) [59], *Π* Model [59], FixMatch [60], USCL [3]) and self-supervised methods (*i.e.*, MoCo v2 [13], SimCLR [10]).

POCUS Pneumonia Detection. On POCUS, we fine-tune the last three layers to verify the transferability of pretrained backbones on the pneumonia detection task. The results are summarized in Table 4. We report the Accuracy of three classes (*i.e.*, COVID-19, pneumonia and the regular), total Accuracy and F1 scores on POCUS. The proposed HiCo achieves the best performance in terms of total Accuracy (*i.e.*, 94.7%) and F1 (94.6%) scores, which are significantly better than other supervised, semi-supervised and self-supervised methods. HiCo also obtains the best Accuracy on COVID-19 and pneumonia classes, and the second-best Accuracy on the regular. USCL achieves the best Accuracy on the regular, which is a semi-supervised CL method with a carefully designed sample pair generation strategy for US video sampling. Compared with USCL, HiCo makes full use of the peer-level and cross-level knowledge from the network via multi-level contrast, which presents a stronger representation capability.

BUSI-BUI Breast Cancer Classification. The fine-tuning results on BUSI-BUI joint dataset are summarized in Table 4. Among the compared methods, HiCo provides the best Accuracy (*i.e.*, 86.0%). Compared with "Supervised", HiCo obtains an absolute gain of 14.7% in terms of Accuracy. We also observe

Table 4. Comparison of fine-tuning results on POCUS and BUSI-BUI classification datasets. Top two results are in bold and underlined.

Method	POCUS					BUSI-BUI
	COVID-19	Pneumonia	Regular	Accuracy (%)	F1 (%)	Accuracy (%)
ImageNet [61]	79.5	78.6	88.6	84.2	81.8	84.9
Supervised [3]	83.7	82.1	86.5	85.0	82.8	71.3
TE [59]	75.7	70.0	89.4	81.7	79.0	71.8
Π Model [59]	77.6	76.4	88.7	83.2	80.6	69.7
FixMatch [60]	83.0	77.5	85.7	83.6	81.6	70.3
MoCo v2 [13]	79.7	81.4	88.9	84.8	82.8	77.8
SimCLR [10]	83.2	89.4	87.1	86.4	86.3	74.6
USCL [3]	90.8	97.0	**95.4**	94.2	94.0	85.5
HiCo (Ours)	**97.1**	**100.0**	92.5	**94.7**	**94.6**	**86.0**

Table 5. Comparison of fine-tuning results on Thyroid US Images dataset. "Supervised", "LFS", and "VCL" denote the backbones pretrained with supervised learning, learning from scratch, and vanilla contrastive learning, respectively.

Method	Precision (%)		Recall (%)		Accuracy (%)
	Benign	Malignant	Benign	Malignant	
Supervised	81.3	89.0	42.6	97.9	88.3
LFS	79.3	88.1	37.7	97.9	87.4
VCL	89.3	88.8	41.0	99.0	88.8
HiCo (Ours)	**91.2**	**89.7**	**49.2**	**99.7**	**90.5**

that our HiCo does not demonstrate significant superiority to "ImageNet" like on POCUS dataset. This is because our pretraining dataset US-4 is captured with convex probes, while BUSI-BUI joint dataset is captured with linear probes. The domain gap between convex and linear probes damages the performance of backbones pretrained on convex probe US data including our HiCo.

4.4 Transferability to Thyroid US Images

We further evaluate the transferability of HiCo on Thyroid US classification dataset. We compare HiCo with the other three backbones pretrained using supervised learning (*i.e.*, Supervised), learning from scratch (*i.e.*, LFS), and vanilla contrastive learning (*i.e.*, VCL). For fair comparisons, all the backbones are pretrained on US-4 dataset and fine-tuned the last three layers. The results are shown in Table 5. We report the Precision and Recall of two classes (*i.e.*, the benign and malignant) and the total Accuracy. We find that HiCo has the consistent best performance on the classification of two classes, and its total Accuracy of 90.5% is also significantly better than the other three methods.

4.5 Cross-modal Transferability to Chest X-ray

The goal of this work is to design an effective method for US video model pretraining. To test our approach's transferability, we also apply our approach to

Table 6. Comparison of fine-tuning results on COVID-Xray-5k chest X-ray dataset. "Supervised", "LFS", and "VCL" denote the backbones pretrained with supervised learning, learning from scratch, and vanilla contrastive learning, respectively.

Method	Precision (%)		Recall (%)		Accuracy (%)
	COVID-19	Normal	COVID-19	Normal	
Supervised	94.2	94.5	90.6	96.7	94.4
LFS	87.2	94.2	90.5	92.1	91.5
VCL	92.2	94.9	91.5	95.4	93.9
HiCo (Ours)	**94.5**	**97.7**	**96.2**	**96.7**	**96.5**

a Chest X-ray classification dataset (*i.e.*, COVID-Xray-5k). This experiment can verify the cross-modal transferability of our approach. The detailed results about Supervised, LFS and VCL are listed in Table 6. From Table 6, we can see that our HiCo achieves the best performance of two classes (*i.e.*, COVID-19 and the normal) and total Accuracy. Specifically, HiCo outperforms the LFS, VCL and Supervised by 5.0%, 2.6% and 2.1%, respectively. Although our approach is designed for US video model pretraining, the above results demonstrate its excellent cross-modal transferability.

5 Conclusion

In this work, we propose the hierarchical contrastive learning for US video model pretraining, which fully and efficiently utilizes both peer-level and cross-level knowledge from a DNN via multi-level contrast, leading to the remarkable transferability for the pretrained model. The advantage of our proposed method is that it flexibly extends the existing CL architecture (*i.e.*, the vanilla contrastive learning framework) and promotes knowledge communication inside the network by designing several simple and effective loss functions instead of designing a complex network structure. We empirically identify that multi-level contrast can greatly accelerate the convergence speed of pretrained models in downstream tasks, and improve the representation ability of models. In addition, a softened objective function is introduced to alleviate the negative effect of some local similarities between different classes, which further facilitates the CL process. Future works include exploiting more general frameworks for multi-level contrast and other applications for US diagnosis.

Acknowledgments. This work is supported by the National Natural Science Foundation of China (No. 62101351), the Guangdong Basic and Applied Basic Research Foundation (No.2020A1515110376), Shenzhen Outstanding Scientific and Technological Innovation Talents PhD Startup Project (No. RCBS202106091 04447108), the National Key R & D Program of China under Grant (No.2021ZD01 13400), Guangdong Provincial Key Laboratory of Big Data Computing, and the Chinese University of Hong Kong, Shenzhen.

References

1. Born, J., Wiedemann, N., Cossio, M., Buhre, C., Brändle, G., Leidermann, K., Goulet, J., Aujayeb, A., Moor, M., Rieck, B., et al.: Accelerating detection of lung pathologies with explainable ultrasound image analysis. Applied Sciences **11** (2021) 672
2. Gao, Y., Maraci, M.A., Noble, J.A.: Describing ultrasound video content using deep convolutional neural networks. In: 2016 IEEE 13th International Symposium on Biomedical Imaging. (2016) 787–790
3. Chen, Y., Zhang, C., Liu, L., Feng, C., Dong, C., Luo, Y., Wan, X.: Uscl: Pretraining deep ultrasound image diagnosis model through video contrastive representation learning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer (2021) 627–637
4. Liu, L., Lei, W., Wan, X., Liu, L., Luo, Y., Feng, C.: Semi-supervised active learning for covid-19 lung ultrasound multi-symptom classification. In: 2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI), IEEE (2020) 1268–1273
5. Gao, L., Zhou, R., Dong, C., Feng, C., Li, Z., Wan, X., Liu, L.: Multi-modal active learning for automatic liver fibrosis diagnosis based on ultrasound shear wave elastography. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), IEEE (2021) 410–414
6. Su, H., Chang, Z., Yu, M., Gao, J., Li, X., Zheng, S., et al.: Convolutional neural network with adaptive inferential framework for skeleton-based action recognition. Journal of Visual Communication and Image Representation **73** (2020) 102925
7. Jiao, J., Droste, R., Drukker, L., Papageorghiou, A.T., Noble, J.A.: Self-supervised representation learning for ultrasound video. In: 2020 IEEE 17th International Symposium on Biomedical Imaging. (2020) 1847–1850
8. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: European conference on computer vision, Springer (2016) 649–666
9. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (2020) 9729–9738
10. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning, PMLR (2020) 1597–1607
11. Jiao, J., Cai, Y., Alsharid, M., Drukker, L., Papageorghiou, A.T., Noble, J.A.: Self-supervised contrastive video-speech representation learning for ultrasound. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer (2020) 534–543
12. Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P.: Contrastive learning of medical visual representations from paired images and text. arXiv preprint arXiv:2010.00747 (2020)
13. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020)
14. Gao, J., Shi, Z., Wang, G., Li, J., Yuan, Y., Ge, S., Zhou, X.: Accurate temporal action proposal generation with relation-aware pyramid network. In: Proceedings of the AAAI Conference on Artificial Intelligence. Volume 34. (2020) 10810–10817
15. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE Trans on Pattern Analysis and Machine Intelligence **39** (2016) 1137–1149

16. Xu, H., Zhang, X., Li, H., Xie, L., Dai, W., Xiong, H., Tian, Q.: Seed the views: Hierarchical semantic alignment for contrastive representation learning. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)
17. Lee, S., Lee, D.B., Hwang, S.J.: Contrastive learning with adversarial perturbations for conditional text generation. 9th International Conference on Learning Representations, ICLR (2021)
18. Li, M., Lin, X., Chen, X., Chang, J., Zhang, Q., Wang, F., Wang, T., Liu, Z., Chu, W., Zhao, D., et al.: Keywords and instances: A hierarchical contrastive learning framework unifying hybrid granularities for text generation. (2022) 4432–4441
19. Li, D., Zhang, T., Hu, N., Wang, C., He, X.: Hiclre: A hierarchical contrastive learning framework for distantly supervised relation extraction. (2022) 2567–2578
20. Wang, X., Wu, Q., Zhang, H., Lyu, C., Jiang, X., Zheng, Z., Lyu, L., Hu, S.: Heloc: Hierarchical contrastive learning of source code representation. arXiv preprint arXiv:2203.14285 (2022)
21. Schmarje, L., Santarossa, M., Schröder, S.M., Koch, R.: A survey on semi-, self-and unsupervised techniques in image classification. arXiv preprint arXiv:2002.08721 (2020)
22. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. (2016) 770–778
23. Gao, J., Sun, X., Ghanem, B., Zhou, X., Ge, S.: Efficient video grounding with which-where reading comprehension. IEEE Transactions on Circuits and Systems for Video Technology (2022)
24. Chi, J., Walia, E., Babyn, P., Wang, J., Groot, G., Eramian, M.: Thyroid nodule classification in ultrasound images by fine-tuning deep convolutional neural network. Journal of digital imaging **30** (2017) 477–486
25. Yap, M.H., Pons, G., Marti, J., Ganau, S., Sentis, M., Zwiggelaar, R., Davison, A.K., Marti, R.: Automated breast ultrasound lesions detection using convolutional neural networks. IEEE journal of biomedical and health informatics **22** (2017) 1218–1226
26. Huang, Q., Luo, Y., Zhang, Q.: Breast ultrasound image segmentation: a survey. International journal of computer assisted radiology and surgery **12** (2017) 493–507
27. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data in Brief **28** (2019)
28. Rodrigues, P.S.: Breast ultrasound image. Mendeley Data, V1, doi: 10.17632/wmy84gzngw.1 (2018)
29. Pedraza, L., Vargas, C., Narváez, F., Durán, O., Muñoz, E., Romero, E.: An open access thyroid ultrasound image database. In: SPIE. Volume 9287. (2015)
30. Nguyen, D.T., Kang, J.K., Pham, T.D., Batchuluun, G., Park, K.R.: Ultrasound image-based diagnosis of malignant thyroid nodule using artificial intelligence. Sensors **20** (2020) 1822
31. Li, P., Zhao, H., Liu, P., Cao, F.: Automated measurement network for accurate segmentation and parameter modification in fetal head ultrasound images. Medical Biological Engineering Computing **58** (2020) 2879–2892
32. Kalafat, E., Yaprak, E., Cinar, G., Varli, B., Ozisik, S., Uzun, C., Azap, A., Koc, A.: Lung ultrasound and computed tomographic findings in pregnant woman with covid-19. Ultrasound in Obstetrics & Gynecology **55** (2020) 835–837
33. Long, L., Zhao, H.T., Zhang, Z.Y., Wang, G.Y., Zhao, H.L.: Lung ultrasound for the diagnosis of pneumonia in adults: a meta-analysis. Medicine **96** (2017)
34. Kerdegari, H., Nhat, P.T.H., McBride, A., Razavi, R., Van Hao, N., Thwaites, L., Yacoub, S., Gomez, A.: Automatic detection of b-lines in lung ultrasound videos

from severe dengue patients. In: 2021 IEEE 18th International Symposium on Biomedical Imaging. (2021) 989–993
35. Wang, X., Burzynski, J.S., Hamilton, J., Rao, P.S., Weitzel, W.F., Bull, J.L.: Quantifying lung ultrasound comets with a convolutional neural network: Initial clinical results. Computers in biology and medicine **107** (2019) 39–46
36. Carrer, L., Donini, E., Marinelli, D., Zanetti, M., Mento, F., Torri, E., Smargiassi, A., Inchingolo, R., Soldati, G., Demi, L., et al.: Automatic pleural line extraction and covid-19 scoring from lung ultrasound data. IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control **67** (2020) 2207–2217
37. Xu, Y., Zhang, Y., Bi, K., Ning, Z., Xu, L., Shen, M., Deng, G., Wang, Y.: Boundary restored network for subpleural pulmonary lesion segmentation on ultrasound images at local and global scales. Journal of Digital Imaging **33** (2020) 1155–1166
38. Razaak, M., Martini, M.G., Savino, K.: A study on quality assessment for medical ultrasound video compressed via hevc. IEEE Journal of Biomedical and Health Informatics **18** (2014) 1552–1559
39. Kwitt, R., Vasconcelos, N., Razzaque, S., Aylward, S.: Localizing target structures in ultrasound video–a phantom study. Medical image analysis **17** (2013) 712–722
40. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention, Springer (2015) 234–241
41. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Semantic image segmentation with deep convolutional nets and fully connected crfs. arXiv preprint arXiv:1412.7062 (2014)
42. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: European conference on computer vision. (2018) 801–818
43. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. arXiv preprint arXiv:2105.05537 (2021)
44. Weng, Y., Zhou, T., Li, Y., Qiu, X.: Nas-unet: Neural architecture search for medical image segmentation. IEEE Access **7** (2019) 44247–44257
45. Huang, R., Ying, Q., Lin, Z., Zheng, Z., Tan, L., Tang, G., Zhang, Q., Luo, M., Yi, X., Liu, P., et al.: Extracting keyframes of breast ultrasound video using deep reinforcement learning. Medical Image Analysis (2022) 102490
46. Gong, B., Shi, J., Han, X., Zhang, H., Huang, Y., Hu, L., Wang, J., Du, J., Shi, J.: Diagnosis of infantile hip dysplasia with b-mode ultrasound via two-stage meta-learning based deep exclusivity regularized machine. IEEE Journal of Biomedical and Health Informatics **26** (2021) 334–344
47. Korbar, B., Tran, D., Torresani, L.: Cooperative learning of audio and video models from self-supervised synchronization. Advances in Neural Information Processing Systems **31** (2018)
48. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv (2018)
49. Ye, M., Zhang, X., Yuen, P.C., Chang, S.F.: Unsupervised embedding learning via invariant and spreading instance feature. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2019) 6210–6219
50. Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving jigsaw puzzles. In: European conference on computer vision, Springer (2016) 69–84
51. Chen, L., Bentley, P., Mori, K., Misawa, K., Fujiwara, M., Rueckert, D.: Self-supervised learning for medical image analysis using image context restoration. Medical image analysis **58** (2019) 101539

52. Doersch, C., Gupta, A., Efros, A.A.: Unsupervised visual representation learning by context prediction. In: Proceedings of the IEEE international conference on computer vision. (2015) 1422–1430
53. Gutmann, M., Hyvärinen, A.: Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In: Proceedings of the thirteenth international conference on artificial intelligence and statistics, JMLR Workshop and Conference Proceedings (2010) 297–304
54. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems **33** (2020) 21271–21284
55. Chen, X., He, K.: Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2021) 15750–15758
56. Chen, X., Xie, S., He, K.: An empirical study of training self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2021) 9640–9649
57. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2017) 2117–2125
58. Minaee, S., Kafieh, R., Sonka, M., Yazdani, S., Soufi, G.J.: Deep-covid: Predicting covid-19 from chest x-ray images using deep transfer learning. Medical image analysis **65** (2020) 101794
59. Laine, S., Aila, T.: Temporal ensembling for semi-supervised learning. In: ICLR. (2017)
60. Sohn, K., Berthelot, D., Li, C., Zhang, Z., Carlini, N., Cubuk, E.D., Kurakin, A., Zhang, H., Raffel, C.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. arXiv:2001.07685 (2020)
61. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision **115** (2015) 211–252