Unsupervised Contrastive Learning of Image Representations from Ultrasound Videos with Hard Negative Mining

Soumen Basu[1] ✉ [0000−0002−3915−7545], Somanshu Singla[1], Mayank Gupta[1],
Pratyaksha Rana[2], Pankaj Gupta[2], and Chetan Arora[1]

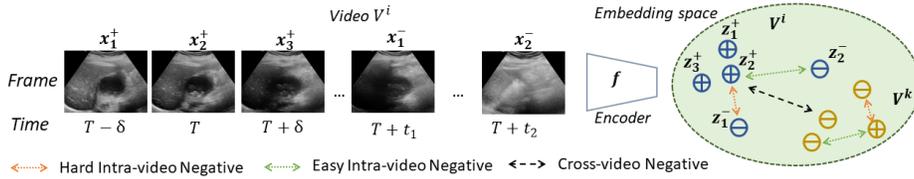[1] Indian Institute of Technology, Delhi, India
soumen.basu@cse.iitd.ac.in
[2] Postgraduate Institute of Medical Education and Research, Chandigarh, India

Abstract. Rich temporal information and variations in viewpoints make video data an attractive choice for learning image representations using unsupervised contrastive learning (UCL) techniques. State-of-the-art (SOTA) contrastive learning techniques consider frames within a video as positives in the embedding space, whereas the frames from other videos are considered negatives. We observe that unlike multiple views of an object in natural scene videos, an Ultrasound (US) video captures different 2D slices of an organ. Hence, there is almost no similarity between the temporally distant frames of even the same US video. In this paper we propose to instead utilize such frames as hard negatives. We advocate mining both intra-video and cross-video negatives in a hardness-sensitive negative mining curriculum in a UCL framework to learn rich image representations. We deploy our framework to learn the representations of Gallbladder (GB) malignancy from US videos. We also construct the first large-scale US video dataset containing 64 videos and 15,800 frames for learning GB representations. We show that the standard ResNet50 backbone trained with our framework improves the accuracy of models pretrained with SOTA UCL techniques as well as supervised pretrained models on ImageNet for the GB malignancy detection task by 2–6%. We further validate the generalizability of our method on a publicly available lung US image dataset of COVID-19 pathologies and show an improvement of 1.5% compared to SOTA. Source code, dataset, and models are available at https://gbc-iitd.github.io/usucl.

Keywords: Contrastive Learning · Ultrasound · Negative Mining

1 Introduction

Due to their remarkable performance, Deep Neural Networks (DNNs) have become defacto standard for a wide range of medical image analysis tasks in recent years [4,6]. However, lack of annotated medical data due to the specialized nature of annotations, and the data privacy issues restrict the applicability of supervised

Fig. 1. We motivate the use of intra-video negatives in contrastive learning. Based on the visibility of a pathology in the intra-video samples, negatives can be sampled. The frames in range $[T - \delta, T + \delta]$ in video V^i has stones and malignant wall thickening visible for a small δ. A slightly distant frame $T + t_1$ shows a GB, but the malignant wall thickening is not visible. This sample acts as a hard negative. The viewing plane further changes in frame $T + t_2$ and GB becomes invisible.

learning of DNNs in medical imaging. Although pretraining on large natural image datasets yields a performance boost for downstream tasks on medical data [3,13], the large domain gap between the natural and medical images remains a bottleneck. Recent works are increasingly going beyond the supervised setup and exploiting unsupervised techniques to compensate for the lack of annotated data [10,19]. Broadly, there are two prominent categories of representation learning techniques for leveraging unlabeled data. In the pretext task-based method, the DNNs are pretrained with some spatial tasks such as image rotation prediction [22] or temporal tasks such as video clip order prediction [26] to learn efficient image representation. On the other hand, contrastive methods [10] try to distinguish between different views of image samples to learn robust representation without labels. Visual representations learned by contrastive learning have been shown to outperform the supervised pretraining on large annotated data in terms of accuracy on the downstream prediction tasks [10,19].

Video data contains rich variations in viewpoints and natural temporal information for objects making it suitable for going beyond image-level contrastive learning. Additionally, the abundance of unlabeled video data makes it an attractive choice for learning representations. Recent works are attempting to exploit the video data to learn robust image-level representations [12,25]. We observe that current SOTA techniques for image representation learning from videos, such as USCL [12] suggest that images coming from the same video are too close to be considered negatives (*similarity conflict*). USCL advocates considering only cross-video samples as negatives to avoid the similarity conflict. While similarity conflict is prevalent for intra-video samples of natural video datasets like action recognition, where each video contains a distinct action, the US videos are inherently different. The frames of a US video contain both types of images where pathology is visible or absent. We argue that frames from the same video where the pathology is absent can be used as hard negatives for the positive samples with the visible pathology. Fig. 1 shows the positive and negative frames from the same US video. The temporal distance between the frames acts as a proxy to the hardness. Negative samples that are temporally closer to the positives contain higher similarities with the positives and are harder to differentiate in

the embedding space. We propose an unsupervised contrastive learning framework to exploit both the intra-video and cross-video negatives for learning robust visual representations from US videos. We design a hardness-sensitive negative mining curriculum to lower the distance between the anchor and negatives gradually. Due to its unsupervised nature, our technique can be used for pretraining a backbone on any US video dataset for superior downstream performance.

We deploy our video contrastive learning framework to pretrain a neural network before supervised fine-tuning to identify GB malignancy in US images. Due to its non-ionizing radiation, low cost, and accessibility, US is a popular non-invasive diagnostic modality for patients with suspected GB pathology. Although there are prior works involving DNNs to detect GB afflictions such as polyp or stones [9,21,23], there is limited prior work on using DNNs to detect GB malignancy in US images [5]. Unlike detecting stones or polyps, identifying GB malignancy from the routine US is challenging for radiologists [16,18]. We observed that ImageNet pretrained classifiers perform even worse than radiologists. We used an in-house abdominal US video dataset to pretrain our model. Our contrastive learning-based pretrained model surpasses human radiologists and SOTA contrastive learning methods on GB malignancy classification.

We also validate our framework on a public lung US dataset, POCUS [8], containing COVID-19 and Pneumonia samples. Pretraining our model on public lung US videos, and finetuning for the downstream classification shows improvement over ImageNet pretraining and the current SOTA contrastive techniques.
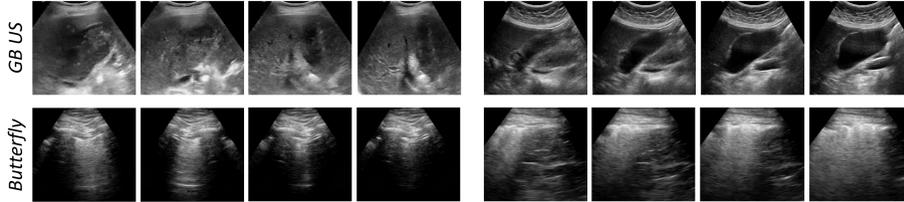**Contributions:** The key contributions of this work are:

- We design an unsupervised contrastive learning technique for US videos to exploit both cross-video and intra-video negatives for rich representation learning. We further use a hardness-sensitive negative mining curriculum to boost the performance of our contrastive learning framework.
- We deploy our framework to solve a novel GB malignancy classification problem from US images. We also validate the efficacy of our technique on a publicly available lung US dataset for COVID detection.
- We are contributing the first US video dataset of 64 videos and 15800 frames containing both malignant and non-malignant GB towards the development of representation learning from medical US videos.

2   Datasets

2.1   In-house US Dataset for Gallbladder Cancer

The transabdominal US video dataset is acquired at the Postgraduate Institute of Medical Education and Research (PGIMER), Chandigarh, India. The PGIMER ethics committee approved the study.
Video Data: Radiologists with 2-8 years of experience in abdominal sonography acquired the data. The US videos were obtained after at least 6 hours of fasting using a 1–5 MHz curved array transducer (C-1-5D, Logiq S8, GE Healthcare).

Fig. 2. Sample video sequences from the GB US video and the Butterfly [1] datasets. Two sequences of size 4 is shown on the left and right for each dataset.

The scanning intended to include the entire GB and the lesion or pathology. The frame rate was 29 fps. The length of the videos varied from 43 to 888 frames depending on the GB distension and size of the lesion. The dataset consists of 32 malignant and 32 non-malignant videos containing a total of 12,251 and 3,549 frames, respectively. Note that we do not use the video level labels in our setup. We cropped the video frames from the center to anonymize the patient information and annotations. The processed video frames were of size 360×480 pixels. Fig. 2 shows some samples from the video data. We are releasing this large scale video dataset for the community.

Image Data: We use the publicly contributed GBCU dataset [5] consisting of 1255 US images from 218 patients. The dataset consists of 990 non-malignant (171 patients with normal and benign GB) and 265 malignant (47 patients) GB images. The images were labeled as normal, benign, or malignant, and such labels were biopsy-proven. We use this dataset for finetuning and report 10-fold cross-validation. Note that the patients recorded in the video dataset are not included in this image dataset to ensure generalization.
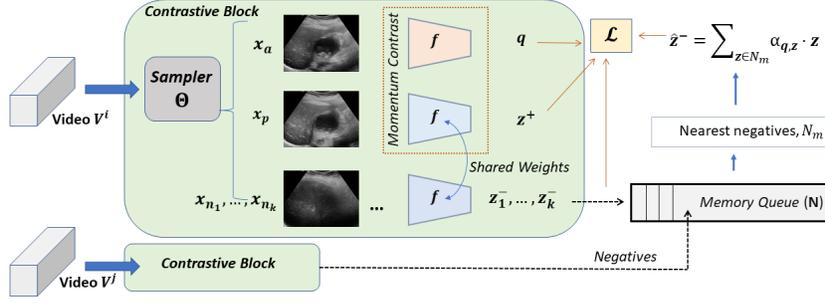
2.2   Public Lung US Dataset for COVID-19

Video Data: We use the public lung US video dataset, Butterfly [1]. Butterfly consists of 22 US videos containing 1533 images of size 658×758 pixels of the lung. The dataset was collected using a Resona 7T machine.

Image Data: We use the publicly available POCUS [8] dataset consisting of a total of 2116 lung US images, of which 655, 349, and 1112 images are of COVID-19, bacterial pneumonia, and healthy control, respectively.

3   Our Method

Contrastive Learning Setup: Suppose, $\mathbf{V}^i = \left\{\mathbf{x}_j\right\}_{j=1}^{M^i}$ is the i-th video in the video dataset consisting of M^i frames where $\mathbf{x}_j$ is the j-th individual frame. We sample an anchor image, $\mathbf{x}_a$, a positive image $\mathbf{x}_p$, and k negative frames $\mathbf{x}_{n_1}, \ldots, \mathbf{x}_{n_k}$ from a video using a sampler $\boldsymbol{\Theta}$. We encode the samples using

Fig. 3. Overview of the proposed contrastive loss, $\mathcal{L}$. An anchor $\mathbf{q}$ and another temporally close sample $\mathbf{z}^+$ from the same video are used as positive pairs. Given the set of cross-video negatives N, and the intra-video negatives $\mathbf{z}_j^-$, we compute $\hat{\mathbf{z}}^-$ from the m most similar cross-video negatives to the anchor. The intra-video samples $\mathbf{z}_j^-$ and $\hat{\mathbf{z}}^-$ are considered as negatives to the anchor.

a backbone f followed by a two layer MLP, g that creates a 128 dimensional embedding vector. We denote the embedding vectors as:

$$\mathbf{q} = g(f(\mathbf{x}_a)), \qquad \mathbf{z}^+ = g(f(\mathbf{x}_p)), \text{ and } \qquad \mathbf{z}_j^- = g(f(\mathbf{x}_{n_j})). \qquad (1)$$

We maximize the agreement between (positive, anchor) and minimize between (anchor, negatives) pairs in the embedding space. Let N be the set of cross video negatives generated from other videos in the dataset. To exploit the cross-video negatives, we calculate the normalized similarity measure between $\mathbf{q}$ and $\mathbf{z}$ for all $\mathbf{z} \in N$:

$$\alpha_{\mathbf{q},\mathbf{z}} = \frac{\exp\left(s(\mathbf{q}, \mathbf{z})/\tau\right)}{\sum_{\mathbf{z}_c \in N} \exp\left(s(\mathbf{q}, \mathbf{z}_c)/\tau\right)}, \qquad (2)$$

where $s(\mathbf{a}, \mathbf{b}) = \mathbf{a} \cdot \mathbf{b}/(||\mathbf{a}||_2 \, ||\mathbf{b}||_2)$ is the cosine similarity and τ is a temperature scaling parameter. We then use the $\alpha_{\mathbf{q},\mathbf{z}}$ to rank the cross-video negatives according to their similarity with the anchor. Note that, with increasing similarity, the hardness of the negatives increase. We pick the top-n hardest cross-video negatives, N_m, and compute $\hat{\mathbf{z}}^- = \sum_{\mathbf{z} \in N_m} \alpha_{\mathbf{q},\mathbf{z}}\mathbf{z}$. Finally, we minimize the loss,

$$\mathcal{L} = -\log \frac{\exp\left(s(\mathbf{q}, \mathbf{z}^+)/\tau\right)}{\exp\left(s(\mathbf{q}, \mathbf{z}^+)/\tau\right) + \sum_{j=1}^{k} \exp\left(s(\mathbf{q}, \mathbf{z}_j^-)/\tau\right) + \exp\left(s(\mathbf{q}, \hat{\mathbf{z}}^-)/\tau\right)}. \qquad (3)$$

Note that $\mathbf{z}_j^-$ is obtained intra-video, and $\hat{\mathbf{z}}^-$ is obtained from inter-video samples. Hence, $\mathcal{L}$ exploits both the intra-video and cross-video hard negatives. We chose $\tau = 0.07$. Value of n was 4 and 2 for GB videos and Butterfly, respectively. **Video Sub-Sampling:** Most SOTA methods sample the anchor and positive pairs uniformly at random from the entire sequence of frames in a video. However, in the case of US videos, the view may change significantly if the samples are

temporally distant. For example, in a transabdominal US video, one sample may show a GB with some parts of a liver while another may show only a liver and not a GB. Pairing such samples as positives would not work for learning a representation of the GB pathology. We recommend sampling the anchor and positives from a temporally close interval. We use a sampler, $\mathbf{\Theta} : \mathbf{V} \to \left(\mathbf{x}_a, \mathbf{x}_p, \{ \mathbf{x}_{n_1}, \ldots, \mathbf{x}_{n_k} \} \right)$ to get the anchor ($\mathbf{x}_a$), positive ($\mathbf{x}_p$), and k negative frames ($\mathbf{x}_{n_1}, \ldots, \mathbf{x}_{n_k}$) from a video $\mathbf{V} = \left\{ \mathbf{x}_j \right\}_{j=1}^{M}$. The indices of the anchor, positive, and negative frames are sampled as following:

$$ a \ \sim \ U\big([1, M]\big), \qquad p \ \sim \ U\big([a - \delta, a + \delta] \setminus \{a\}\big), $$

$$ n_1, \ldots, n_k \ \overset{\text{i.i.d.}}{\sim} \ U\big([1, M] \setminus [a - \Delta, a + \Delta]\big) $$

where $U(I)$ denotes sampling uniformly at random from interval I. We vary Δ between the Δ_h and Δ_l during the curriculum to adjust the hardness of the mined negatives. Also, $1 \leq \delta \ll M$ and $\Delta_l \leq \Delta \leq \Delta_h < M$.

Curriculum-based Negative Mining: We use the negative samples in a hardness-sensitive order for effective learning. The model would initially learn to distinguish anchors from distant negatives and then gradually closer and thus harder negatives will be introduced. We start the training with only cross-video negatives and minimize the loss term,

$$ \mathcal{L}_{\text{cross}} = - \log \frac{\exp\big(s(\mathbf{q}, \mathbf{z}^+)/\tau\big)}{\exp\big(s(\mathbf{q}, \mathbf{z}^+)/\tau\big) + \exp\big(s(\mathbf{q}, \hat{\mathbf{z}}^-)/\tau\big)}. \tag{4} $$

We then gradually start using the loss in Eq. (3) to introduce intra-video negatives, which are more challenging to distinguish from the anchor than the cross-video negatives. We initially keep the $\Delta = \Delta_h = \lceil M/5 \rceil$ used in Eq. (4) for sampling the negatives and ensure the anchor and negatives are at least Δ frames apart temporally, and the hardness is comparatively lower. We gradually lower the Δ using a cosine annealing during the later phase of training to introduce harder negatives. We chose $\delta = 3$, $k = 3$, and $\Delta_l = 7$ in our experiments.

4   Experiments and Results

Experimental Setup: We use a machine with Intel Xeon Gold 5218@2.30GHz processor and 4 Nvidia Tesla V100 GPUs for our experiments. We pretrain a ResNet50 encoder using SGD with LR 0.003, weight decay 10^{-4}, and momentum 0.9 for 60 epochs with batch size 32. We used a grid-search strategy to select the sampling hyper-parameters. We use a cosine annealing of the LR. The parameters of the anchor and positive encoders are updated using momentum contrast with momentum coefficient, $m = 0.999$. Size of the queue for cross-video negative set is $|N| = 96$ for GB videos and $|N| = 66$ for Butterfly. We fine-tune for 30 epochs with batch size of 64 and an SGD optimizer with weight decay $5 \cdot 10^{-4}$. The remaining hyper-parameters are set similar to that of the pretraining phase.

Comparison with SOTA: We compare the ResNet50 [20] backbone pretrained on our contrastive learning framework with ImageNet pretraining, SOTA UCL

Table 1. The fine-tuning performance of ResNet50 model in classifying malignant vs. non-malignant GBs from US images. We report accuracy, specificity, and sensitivity.

Method	Acc.	Spec.	Sens.
Pretrained on [14]	0.867 ± 0.070	0.926 ± 0.069	0.672 ± 0.147
SimCLR [10]	0.897 ± 0.040	0.912 ± 0.055	0.874 ± 0.067
SimSiam [11]	0.900 ± 0.052	0.913 ± 0.059	0.861 ± 0.061
BYOL [15]	0.844 ± 0.129	0.871 ± 0.144	0.739 ± 0.178
MoCo v2[19]	0.886 ± 0.061	0.893 ± 0.078	0.871 ± 0.094
Cycle-Contrast [25]	0.861 ± 0.087	0.867 ± 0.098	0.844 ± 0.097
USCL [12]	0.901 ± 0.047	0.923 ± 0.041	0.831 ± 0.072
Ours	$\mathbf{0.921 \pm 0.034}$	$\mathbf{0.926 \pm 0.043}$	$\mathbf{0.900 \pm 0.046}$

Table 2. Comparison of finetuning performance of ResNet using the SOTA USCL, ImageNet pretraining, and our method on POCUS. We used the official finetuning script used by USCL. The pretraining was done on Butterfly dataset. The USCL official script reports the average accuracy over 5 runs. **C**, **P**, and **R** denote COVID-19, Pneumonia, and Regular respectively.

Table 3. We asked expert radiologists to classify GB malignancy for the test set of GBCU containing 80 non-malignant, and 42 malignant GB US images. Radiologists were not allowed access to any other patient data. The performance of the expert radiologists is comparable to that reported in the literature [7,17].

Method	Accuracy			
	Overall	C	P	R
Pretrained [14]	0.842	0.795	0.786	0.886
SimCLR	0.864	0.832	0.894	0.871
MoCo v2	0.848	0.797	0.814	0.889
USCL	0.907	0.861	0.903	0.935
Ours	**0.922**	**0.892**	**0.951**	0.931

Method	Acc.	Spec.	Sens.
Radiologist A	0.816	0.873	0.707
Radiologist B	0.784	0.811	0.732
USCL	0.812	0.838	0.762
Pretrained [14]	0.787	0.875	0.619
Ours	0.877	0.900	0.833

techniques SimCLR [10], SimSiam [11], MoCo [19], BYOL [15], and SOTA image representation learning from video methods: Cycle-Contrast [25] and USCL [12]. USCL is specialized for pretraining on the US datasets. We note the performance of our pretraining framework for the GB malignancy classification task in Tab. 1. Our method gives 92.1% overall accuracy which is 2% higher than SOTA and 90% accuracy on malignant samples (sensitivity) which is 7% higher than USCL. Our method also outperforms human experts significantly for detecting GB malignancy from US images (Tab. 3). We show the performance comparison on the POCUS dataset in Tab. 2 and observe that our method surpasses the SOTA USCL pretraining by 1.5%. Fig. S1 in the supplementary material shows the Grad-CAM [24] visuals of the last conv layer to demonstrate that the attention regions of contrastive backbones are more precise and clinically relevant.

Generality of our Method: We have shown our method's efficacy on two different tasks - (1) GB malignancy detection from abdominal US and (2) COVID-

Table 4. Significance of joint mining of intra, and cross video negatives. While the individual mining techniques match SOTA performance in GB malignancy, pretraining with proposed joint mining surpasses the current SOTA.
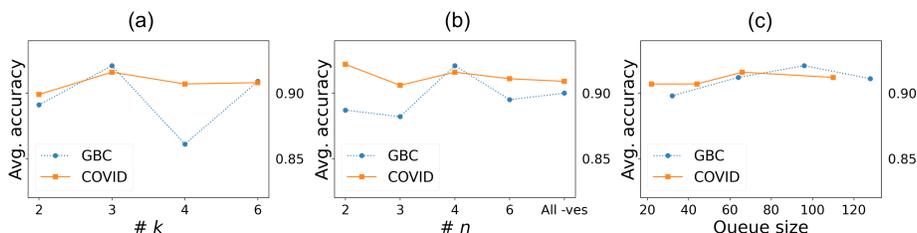
Type of negative used		Acc.	Spec.	Sens.
Cross-video	Intra-video			
✓		0.890 ± 0.062	0.897 ± 0.061	0.869 ± 0.108
	✓	0.893 ± 0.057	0.904 ± 0.059	0.835 ± 0.109
✓	✓	0.921 ± 0.034	0.926 ± 0.043	0.900 ± 0.046

Table 5. Effectiveness of our curriculum-based negative mining. The other alternative, curricula based trained models, lag significantly in GB malignancy classification.

Method	Acc.	Spec.	Sens.
Proposed curriculum	0.921 ± 0.034	0.926 ± 0.043	0.900 ± 0.046
Anti-curriculum	0.887 ± 0.064	0.902 ± 0.056	0.836 ± 0.097
Control-curriculum	0.897 ± 0.067	0.918 ± 0.062	0.810 ± 0.114

19 detection from lung US, which establishes the generality of our method on US modality. We also performed preliminary analysis on the performance of a ResNet50 classifier in detecting COVID-19 from a public CT dataset [27]. We pretrained the model on another CT dataset [2]. The (accuracy, specificity, sensitivity) for our method was (0.80, 0.81, 0.80) as compared to (0.73, 0.72, 0.74) of ImageNet pretraining, and (0.78, 0.81, 0.76) of USCL. The results are indicative of the generality of our method across modalities.

Ablation Study: (1) **Significance of Mining Intra-Video Hard Negatives:** In Tab. 4 we observe that when the intra-video negatives are not used, the performance of malignancy detection of our method becomes comparable to that of Cycle-Contrast and USCL; both methods use only cross-video negatives. On the other hand, if only intra-video negatives are used, the model performance becomes similar to that of the SOTA image contrastive techniques. This shows the importance of mining both intra-video and cross-video negatives in achieving the performance boost. (2) **Effectiveness of Curriculum-based Negative Mining:** We propose to use increasing order of hardness for negative mining during the training. To assert the effectiveness of such a curriculum, we compare the curriculum with two possible alternatives - (i) *anti-curriculum* initially trains with harder negatives and progressively lowers the hardness of the negatives, and (ii) *control-curriculum* does not order the negatives according to their hardness. We initially train the model with temporally close intra-video negatives during the anti-curriculum and gradually start sampling temporally distant negatives. During the last few epochs, only the cross-video negatives are used. Tab. 5 shows the performance comparison of the proposed hardness-sensitive curriculum over the two alternative curricula. (3) **Sensitivity of Hyper-parameters:** Fig. 4 shows the sensitivity of three important pretraining hyper-parameters on the accuracy of the downstream task for our method.

Fig. 4. Sensitivity of pretraining hyper-parameters - (a) number of intra-video nega-tives (k), (b) number of top cross-video negatives (n), and (c) queue size ($|N|$) - on downstream accuracy. Mean cross-val accuracy is shown for GB Cancer and COVID.

5   Conclusion

We introduce the first large-scale US video dataset for learning GB malig-nancy representation and propose an efficient UCL framework that exploits both intra-video and cross-video negatives through a hardness-aware curriculum. Our framework surpasses the human experts, imageNet-pretrained DNNs, and DNNs pretrained with SOTA contrastive learning methods specialized for US modality.

References

1. Butterfly videos. https://www.butterflynetwork.com/index.html, accessed: 2022-03-02 4
2. Afshar, P., Heidarian, S., Enshaei, N., Naderkhani, F., Rafiee, M.J., Oikonomou, A., Fard, F.B., Samimi, K., Plataniotis, K.N., Mohammadi, A.: Covid-ct-md, covid-19 computed tomography scan dataset applicable in machine learning and deep learning. Scientific Data **8**(1), 1–8 (2021) 8
3. Alzubaidi, L., Fadhel, M.A., Al-Shamma, O., Zhang, J., Santamaría, J., Duan, Y., Oleiwi, S.R.: Towards a better understanding of transfer learning for medical imaging: a case study. Applied Sciences **10**(13), 4523 (2020) 2
4. Ardila, D., Kiraly, A.P., Bharadwaj, S., Choi, B., Reicher, J.J., Peng, L., Tse, D., Etemadi, M., Ye, W., Corrado, G., et al.: End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nature medicine **25**(6), 954–961 (2019) 1
5. Basu, S., Gupta, M., Rana, P., Gupta, P., Arora, C.: Surpassing the human ac-curacy: Detecting gallbladder cancer from usg images with curriculum learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20886–20896 (2022) 3, 4
6. Bejnordi, B.E., Veta, M., Van Diest, P.J., Van Ginneken, B., Karssemeijer, N., Litjens, G., Van Der Laak, J.A., Hermsen, M., Manson, Q.F., Balkenhol, M., et al.: Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. Jama **318**(22), 2199–2210 (2017) 1
7. Bo, X., Chen, E., Wang, J., Nan, L., Xin, Y., Wang, C., Lu, Q., Rao, S., Pang, L., Li, M., et al.: Diagnostic accuracy of imaging modalities in differentiating xan-thogranulomatous cholecystitis from gallbladder cancer. ATM **7**(22) (2019) 7
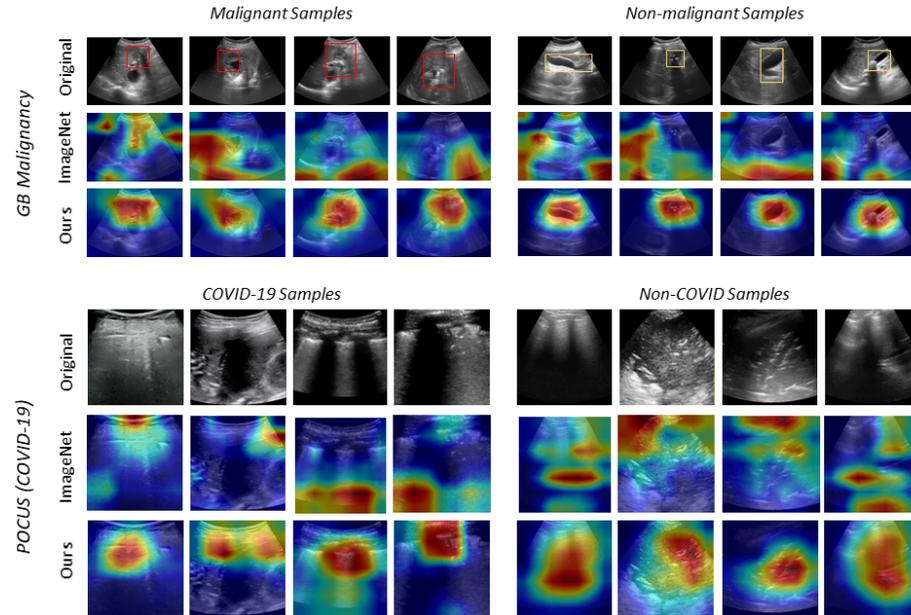
8. Born, J., Brändle, G., Cossio, M., Disdier, M., Goulet, J., Roulin, J., Wiedemann, N.: Pocovid-net: automatic detection of covid-19 from a new lung ultrasound imaging dataset (pocus). arXiv preprint arXiv:2004.12084 (2020) 3, 4

9. Chen, T., Tu, S., Wang, H., Liu, X., Li, F., Jin, W., Liang, X., Zhang, X., Wang, J.: Computer-aided diagnosis of gallbladder polyps based on high resolution ultrasonography. Computer methods and programs in biomedicine **185**, 105118 (2020) 3

10. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML. pp. 1597–1607. PMLR (2020) 2, 7

11. Chen, X., He, K.: Exploring simple siamese representation learning. In: CVPR. pp. 15750–15758 (2021) 7

12. 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: MICCAI. pp. 627–637. Springer (2021) 2, 7

13. Cheng, P.M., Malhi, H.S.: Transfer learning with convolutional neural networks for classification of abdominal ultrasound images. Journal of digital imaging **30**(2), 234–243 (2017) 2

14. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR. pp. 248–255 (2009) 7

15. 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. NIPS **33**, 21271–21284 (2020) 7

16. Gupta, P., Dutta, U., Rana, P., Singhal, M., Gulati, A., Kalra, N., Soundararajan, R., Kalage, D., Chhabra, M., Sharma, V., et al.: Gallbladder reporting and data system (gb-rads) for risk stratification of gallbladder wall thickening on ultrasonography: an international expert consensus. Abdominal Radiology pp. 1–12 (2021) 3

17. Gupta, P., Kumar, M., Sharma, V., Dutta, U., Sandhu, M.S.: Evaluation of gallbladder wall thickening: a multimodality imaging approach. Expert Rev. Gastroenterol. Hepatol. **14**(6), 463–473 (2020) 7

18. Gupta, P., Marodia, Y., Bansal, A., Kalra, N., Kumar-M, P., Sharma, V., Dutta, U., Sandhu, M.S.: Imaging-based algorithmic approach to gallbladder wall thickening. World journal of gastroenterology **26**(40), 6163 (2020) 3

19. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR. pp. 9729–9738 (2020) 2, 7

20. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016) 6

21. Jeong, Y., Kim, J.H., Chae, H.D., Park, S.J., Bae, J.S., Joo, I., Han, J.K.: Deep learning-based decision support system for the diagnosis of neoplastic gallbladder polyps on ultrasonography: preliminary results. Scientific Reports **10**(1), 1–10 (2020) 3

22. Komodakis, N., Gidaris, S.: Unsupervised representation learning by predicting image rotations. In: International Conference on Learning Representations (ICLR) (2018) 2

23. Lian, J., Ma, Y., Ma, Y., Shi, B., Liu, J., Yang, Z., Guo, Y.: Automatic gallbladder and gallstone regions segmentation in ultrasound image. International journal of computer assisted radiology and surgery **12**(4), 553–568 (2017) 3

24. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Gradcam: Visual explanations from deep networks via gradient-based localization. In:

Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017) 7

25. Wu, H., Wang, X.: Contrastive learning of image representations with cross-video cycle-consistency. In: ICCV. pp. 10149–10159 (2021) 2, 7

26. Xu, D., Xiao, J., Zhao, Z., Shao, J., Xie, D., Zhuang, Y.: Self-supervised spatiotemporal learning via video clip order prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10334–10343 (2019) 2

27. Yang, X., He, X., Zhao, J., Zhang, Y., Zhang, S., Xie, P.: Covid-ct-dataset: a ct scan dataset about covid-19. arXiv preprint arXiv:2003.13865 (2020) 8

Supplementary Material

A   Visualization of Feature Representation

Fig. S1 shows the Grad-CAM visuals using the features generated by the last convolutional layer.

Fig. S1. Grad-CAM visuals of the last conv layer in ImageNet pretrained model and the model pretrained using our CL method. The attention regions of contrastive backbones are more precise and clinically relevant.

Table S1. Finetuning performance of ResNet18 for classifying malignant vs. non-malignant GBs from USG images. Both ResNet50 (result in main text) and ResNet18 backbones show better accuracy and sensitivity of GB malignancy detection with contrastive pretraining as compared to the ImageNet pretraining.

Method	Acc.	Spec.	Sens.
Pretrained on ImageNet	0.844 ± 0.053	0.856 ± 0.054	0.795 ± 0.097
USCL	0.896 ± 0.061	0.916 ± 0.066	0.833 ± 0.099
Ours	0.907 ± 0.064	0.919 ± 0.072	0.862 ± 0.069

Table S2. Comparison of finetuning performance of ResNet18 using the SOTA USCL, ImageNet pretraining, and our method on POCUS. The pretraining was done on Butterfly dataset.

Method	Accuracy			
	Overall	COVID-19	Pneumonia	Regular
Pretrained on ImageNet	0.874	0.797	0.874	0.919
USCL	0.914	0.916	0.940	0.906
Ours	0.922	0.902	0.946	0.926

B   Performance with Another Backbone

We compare the fine-tuning performance of ResNet18 backbone by pretraining with our method is compared with ImageNet-pretrained ResNet18 and USCL-pretrained ResNet18. We show the results in Tabs. S1 and S2. In the main text, we showed performance analysis with ResNet50 backbone. The superior performance of both backbones pretrained with our method indicates the generalizability of our framework to multiple backbones.