Intra-video Positive Pairs in Self-Supervised Learning for Ultrasound

Anonymous Author(s)
Affiliation
Address
email

Abstract

The videographic nature of ultrasound offers flexibility for defining the similarity relationship between pairs of images for self-supervised learning (SSL). In this study, we investigated the effect of utilizing proximal, distinct images from the same ultrasound video as pairs for joint embedding SSL. Additionally, we introduced a sample weighting scheme that increases the weight of closer image pairs and demonstrated how it can be integrated into SSL objectives. Named *Intra-Video Positive Pairs* (IVPP), the method surpassed previous ultrasound-specific contrastive learning methods' average test accuracy on COVID-19 classification with the POCUS dataset by $\geq 1.3\%$. Investigations revealed that some combinations of IVPP hyperparameters can lead to improved or worsened performance, depending on the downstream task.

1 Introduction

Deep learning has been extensively studied as a means to automate diagnostic tasks in medical ultrasound (US) [1–5]. Barriers to the development of such systems include lack of open access to large datasets, along with the cost and expertise required to label vast amounts of institutionally acquired examinations. Additionally, many US examinations in acute care are not archived or documented [6, 7]. Joint embedding self-supervised learning (SSL) has been explored as a means to leverage unlabelled data for representation learning with US [8–11]. A common way to define positive pairs for SSL is to apply two stochastic transformations to an image, producing two distorted views with similar content. Being a video-based examination, US offers a unique opportunity to compose alternative pairwise relationships. However, due to the dynamic nature of US, all frames in a US video may not possess the same label for all downstream US interpretation tasks. Moreover, US videos from the same examination or patient may bear a striking resemblance to each other.

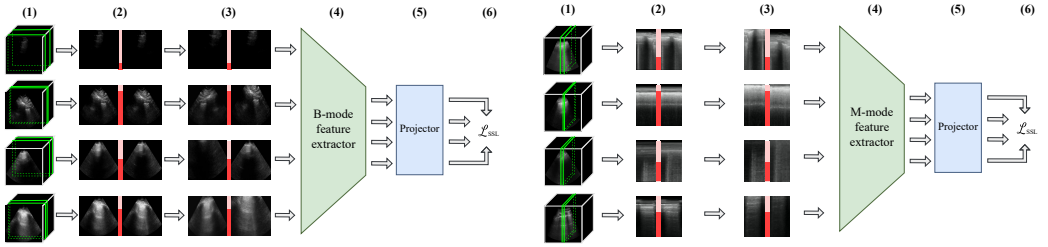
In this study, we examined the effect of proximity and sample weighting of intra-video positive pairs for common SSL methods applied to US. We evaluated on two tasks: one for B-mode US (i.e., US video composed from several scan lines) and M-mode US (i.e., images depicting the evolution of one US scan line over time). Our contributions are summarized as follows:

- A method for sampling US intra-video positive pairs for joint embedding SSL.
- A sample weighting scheme for joint embedding SSL methods that weighs positive pairs according to the temporal or spatial distance between them in their video of origin
- A comprehensive assessment of intra-video positive pairs integrated with contrastive and non-contrastive SSL methods, as measured by downstream performance in B-mode and M-mode lung US classification tasks.

Figure 1 summarizes the methods proposed in this study. Although previous studies have formulated contrastive learning methods with intra-video positive pairs for US [8, 10, 12, 13], the authors believe

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

there are no preceding studies that investigated the effect of sampling multiple images from the same US video in non-contrastive learning. More generally, we believe that this study is the first to integrate sample weights into non-contrastive objectives.



(a) For B-mode ultrasound, positive pairs are temporally separated images from the same video.

(b) For M-mode ultrasound, positive pairs are spatially separated images from the same video.

Figure 1: An overview of the methods introduced in this study. Positive pairs of images separated by no more than a threshold are sampled from the same B-mode video **(1)**. Sample weights inversely proportional to the separation between each image (red bars) are calculated for each pair **(2)**. Random transformations are applied to each image **(3)**. Images are sent to a neural network consisting of a feature extractor **(4)** and a projector **(5)** connected in series. The outputs are used to calculate the self-supervised objective $\mathcal{L}_{\text{SSL}}$ **(6)**.

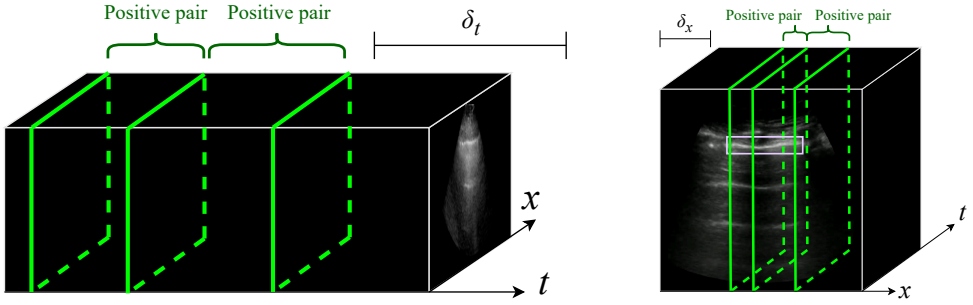
2 Methods

Datasets: As done in previous studies on on US-specific joint embedding methods [8, 10, 12, 13], we evaluate on the public POCUS lung US dataset [14]. This dataset contains 140 publicly sourced US videos (2116 images) labelled for three classes: COVID-19 pneumonia, non-COVID-19 pneumonia, and normal lung. Pretraining is conducted on the public Butterfly dataset, which contains 22 unlabelled lung ultrasound videos [15]. We also utilize a private dataset of 25 917 parenchymal lung US videos (5.9×10^6 images), hereafter referred to as *ParenchymalLUS*. Of these videos, 20 000 had no labels. We evaluated on two binary classification tasks: A-lines versus B-line classification (i.e., `AB`), and lung sliding classification (i.e., `LS`). Details on ParechymalLUS and descriptions of the downstream tasks can be found in Appendix A.

Intra-video Positive Pairs (IVPP): Clinically relevant patterns commonly surface and disappear within the same US video as the US probe and/or the patient move. Without further knowledge of the US examinations in an unlabelled dataset, we conjectured that it may be safest to only assume that positive pairs are intra-video images that are close to each other. Our method distinguishes itself from prior work by only considering proximal frames to be positive pairs and disregarding distant intra-video pairs, treating them as neither positive nor negative pairs.

For B-mode US videos, we define positive pairs as intra-video images $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ that are separated by no more than δ_{t} seconds (Figure 2a). To accomplish this, $\boldsymbol{x}_1$ is randomly selected from the video's images, and $\boldsymbol{x}_2$ is randomly drawn from the set of images within δ_{t} seconds of $\boldsymbol{x}_1$. Videos with higher frame rates provide more positive pair candidates, potentially increasing the diversity of pairs due to naturally occurring noise. A similar sampling scheme is applied for M-mode US images. Like previous studies, we define M-mode images as vertical slices through the time axis of a B-mode video, taken at a specific x-coordinate [11, 16, 17]. Accordingly, M-mode images are columns of B-mode pixels for every frame, concatenated horizontally. We define positive pairs as M-mode images whose x-coordinates differ by no more than δ_{x} pixels (Figure 2b). To avoid resolution differences, B-mode videos are resized to the same dimensions prior to sampling M-mode images.

Sample Weights: The chance that intra-video images are semantically related increases as temporal or spatial separation decreases. To temper the effect of unrelated positive pairs, we applied sample weights to positive pairs in the SSL objective according to their temporal or spatial distance. Sample weights were incorporated into each SSL objective trialled in this study: SimCLR [18], Barlow Twins [19], and VICReg [20]. Appendix B details how sample weights are calculated and integrated into the SSL objectives.



(a) B-mode (b) M-mode

Figure 2: For B-mode ultrasound, positive pairs are frames in the same video that are within δ_t seconds of each other. For M-mode ultrasound, positive pairs are M-mode images originating from the same B-video that are no more than δ_x pixels apart. In the context of lung ultrasound, M-mode images should intersect the pleural line (outlined in mauve).

3 Experiments & Evaluation

Fine-tuning Performance: We evaluated our approach on each of the downstream tasks and report the mean cross-validation accuracy for COVID, along with the test set AUC for AB and LS. Pretraining and training protocols can be found in Appendix C. Multiple values of the threshold parameter were investigated, with and without sample weights. For the COVID and AB tasks, we examined $\delta_t \in \{0, 0.5, 1, 1.5\}$ seconds. For LS we explored $\delta_x \in \{0, 5, 10, 15\}$ pixels. As shown in Figure 3, mean cross-validation accuracy of each fine-tuned model peaked at nonzero values of δ_t. Sample weights decreased performance when $\delta = 0.5$, but were helpful when $\delta = 1.0$ and $\delta = 1.5$. Figure 4 gives fine-tuning performance for AB and LS. We observed no discernible trend for the effect of sample weights that was consistent for any task, pretraining method, δ_t, or δ_x. A striking finding across AB and LS was that SimCLR consistently outperformed Barlow Twins and VICReg, which are both non-contrastive methods. Evaluation via linear classification was also conducted, revealing similar results (see Appendix D).

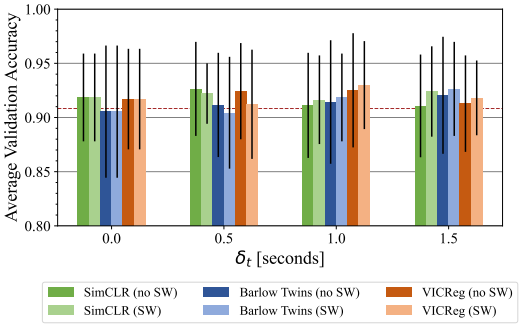
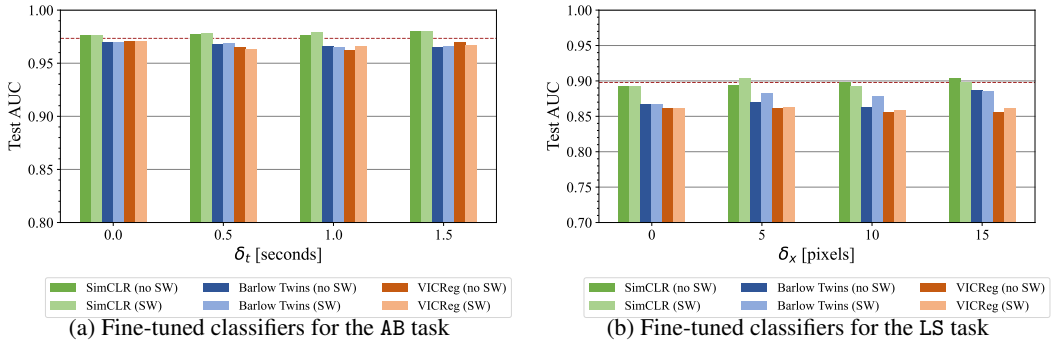


Figure 3: Average test accuracy across 5-fold cross validation on the POCUS dataset, pretrained using a variety of values for δ_t, with and without sample weights. The dashed line indicates initialization with ImageNet-pretrained weights.

Comparison to Baselines: We compared IVPP (with the best hyperparameter assignments) for each SSL objective against USCL [8] and UCL [10], which are preexisting US contrastive learning methods. As shown in Table 1, IVPP outperformed all baseline methods on POCUS, regardless of the pretraining objective. USCL and IVPP with SimCLR performed comparably on the AB task. On the LS task, which is more fine-grained and had a stronger class imbalance, IVPP with SimCLR achieved the greatest performance. Non-contrastive methods were unremarkable, achieving lower performance than networks initialized with ImageNet-pretrained weights.

Label Efficiency: We devised an experiment to compare the performance of models pretrained using different IVPP hyperparameters in low-label settings. The ParenchymalLUS training set was split by patient identifier into 20 subsets. Pretrained models were fine-tuned on each subset, resulting in a population of 20 test set metrics for each hyperparameter combination. As is visible in Figure 5, SimCLR obtained the best performance by a large margin. As detailed in Appendix D, the differences between some means were statistically significant.

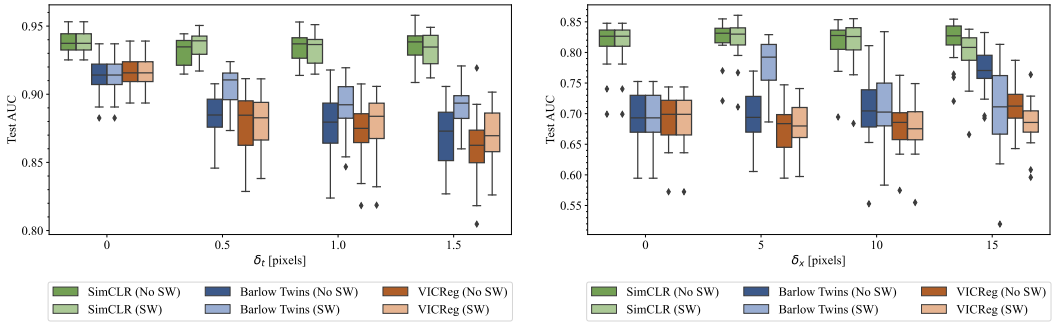


(a) Fine-tuned classifiers for the AB task (b) Fine-tuned classifiers for the LS task

Figure 4: ParenchymalLUS test set AUC for models fine-tuned for the AB and LS binary classification tasks and pretrained with a variety of intra-video positive pair thresholds, with and without sample weights (SW). The dashed line indicates initialization with ImageNet-pretrained weights.

Dataset	POCUS	ParenchymalLUS	
Task	COVID Mean (std) test accuracy	AB Test AUC	LS Test AUC
Random initialization	0.881 (0.050)	0.954	0.790
ImageNet initialization	0.908 (0.043)	0.973	0.898
USCL [8]	0.905 (0.044)	0.979	0.874
US UCL [10]	0.901 (0.054)	0.967	0.809
IVPP [SimCLR]	0.926 (0.043)	**0.980**	**0.903**
IVPP [Barlow Twins]	0.921 (0.054)	0.969	0.887
IVPP [VICReg]	**0.930** (**0.046**)	0.971	0.862

Table 1: Performance of fine-tuned models pretrained using IVPP compared to US-specific contrastive learning methods, USCL and UCL, and to baseline Random and ImageNet initializations.



(a) AB task (b) LS task

Figure 5: Distributions of test AUC for each pretraining method and assignment to δ, with and without sample weights. Each experiment is repeated 20 times on disjoint subsets of the training set, each containing all images from a group of patients.

4 Discussion

Overall, the results indicated that the optimal assignment for IVPP hyperparameters may be problem-specific. First, IVPP may improve performance on downstream ultrasound interpretation tasks; however, practitioners are advised to include a range of values of δ with and without sample weights in their hyperparameter search. Subsequent work could explore IVPP for other downstream tasks in US outside of the lung. Second, SimCLR outperformed non-contrastive methods across multiple tasks – contrary to our initial belief. Future work assessing non-contrastive methods for tasks in US examinations or alternative imaging modalities would shed light on the utility of non-contrastive methods outside the typical evaluation setting of photographic images.

References

[1] M. R. Whitson and P. H. Mayo, "Ultrasonography in the emergency department," *Critical Care*, vol. 20, pp. 1–8, 2016.

[2] Y. H. Lau and K. C. See, "Point-of-care ultrasound for critically-ill patients: A mini-review of key diagnostic features and protocols," *World Journal of Critical Care Medicine*, vol. 11, no. 2, p. 70, 2022.

[3] N. J. Soni, R. Arntfield, and P. Kory, *Point-of-Care Ultrasound*. Philadelphia: Elsevier, second ed., 2020.

[4] R. Sood, A. F. Rositch, D. Shakoor, E. Ambinder, K.-L. Pool, E. Pollack, D. J. Mollura, L. A. Mullen, and S. C. Harvey, "Ultrasound for breast cancer detection globally: a systematic review and meta-analysis," *Journal of global oncology*, vol. 5, pp. 1–17, 2019.

[5] E. S. Yim and G. Corrado, "Ultrasound in sports medicine: relevance of emerging techniques to clinical care of athletes," *Sports medicine*, vol. 42, pp. 665–680, 2012.

[6] M. K. Hall, J. Hall, C. P. Gross, N. J. Harish, R. Liu, S. Maroongroge, C. L. Moore, C. C. Raio, and R. A. Taylor, "Use of point-of-care ultrasound in the emergency department: insights from the 2012 medicare national payment data set," *Journal of Ultrasound in Medicine*, vol. 35, no. 11, pp. 2467–2474, 2016.

[7] R. Kessler, J. R. Stowell, J. A. Vogel, M. M. Liao, and J. L. Kendall, "Effect of interventional program on the utilization of pacs in point-of-care ultrasound," *Journal of digital imaging*, vol. 29, pp. 701–705, 2016.

[8] Y. Chen, C. Zhang, L. Liu, C. Feng, C. Dong, Y. Luo, and X. Wan, "USCL: Pretraining Deep Ultrasound Image Diagnosis Model Through Video Contrastive Representation Learning," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24*, pp. 627–637, Springer, 2021.

[9] D. Anand, P. Annangi, and P. Sudhakar, "Benchmarking self-supervised representation learning from a million cardiac ultrasound images," in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 529–532, IEEE, 2022.

[10] S. Basu, S. Singla, M. Gupta, P. Rana, P. Gupta, and C. Arora, "Unsupervised contrastive learning of image representations from ultrasound videos with hard negative mining," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 423–433, Springer, 2022.

[11] B. VanBerlo, B. Li, A. Wong, J. Hoey, and R. Arntfield, "Exploring the utility of self-supervised pretraining strategies for the detection of absent lung sliding in m-mode lung ultrasound," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3076–3085, 2023.

[12] Y. Chen, C. Zhang, C. H. Ding, and L. Liu, "Generating and weighting semantically consistent sample pairs for ultrasound contrastive learning," *IEEE Transactions on Medical Imaging*, 2022.

[13] C. Zhang, Y. Chen, L. Liu, Q. Liu, and X. Zhou, "Hico: Hierarchical contrastive learning for ultrasound video model pretraining," in *Proceedings of the Asian Conference on Computer Vision*, pp. 229–246, 2022.

[14] J. Born, G. Brändle, M. Cossio, M. Disdier, J. Goulet, J. Roulin, and N. Wiedemann, "POCOVID-Net: Automatic Detection of COVID-19 From a New Lung Ultrasound Imaging Dataset (POCUS)," *arXiv preprint arXiv:2004.12084*, 2020.

[15] Butterfly Network, "Covid-19 ultrasound gallery." `https://www.butterflynetwork.com/covid19/covid-19-ultrasound-gallery`, 2020. Accessed: September 20, 2020.

[16] M. Jaščur, M. Bundzel, M. Malík, A. Dzian, N. Ferenčík, and F. Babič, "Detecting the absence of lung sliding in lung ultrasounds using deep learning," *Applied Sciences*, vol. 11, no. 15, p. 6976, 2021.

[17] B. VanBerlo, D. Wu, B. Li, M. A. Rahman, G. Hogg, B. VanBerlo, J. Tschirhart, A. Ford, J. Ho, J. McCauley, *et al.*, "Accurate assessment of the lung sliding artefact on lung ultrasonography using a deep learning approach," *Computers in Biology and Medicine*, vol. 148, p. 105953, 2022.

[18] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*, pp. 1597–1607, PMLR, 2020.

[19] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *International Conference on Machine Learning*, pp. 12310–12320, 2021.

[20] A. Bardes, J. Ponce, and Y. LeCun, "VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning," in *International Conference on Learning Representations*, 2022.

[21] R. Arntfield, D. Wu, J. Tschirhart, B. VanBerlo, A. Ford, J. Ho, J. McCauley, B. Wu, J. Deglint, R. Chaudhary, *et al.*, "Automation of Lung Ultrasound Interpretation via Deep Learning for the Classification of Normal Versus Abnormal Lung Parenchyma: A Multicenter Study," *Diagnostics*, vol. 11, no. 11, p. 2049, 2021.

[22] B. VanBerlo, D. Smith, J. Tschirhart, B. VanBerlo, D. Wu, A. Ford, J. McCauley, B. Wu, R. Chaudhary, C. Dave, *et al.*, "Enhancing annotation efficiency with machine learning: Automated partitioning of a lung ultrasound dataset by view," *Diagnostics*, vol. 12, no. 10, p. 2351, 2022.

[23] B. VanBerlo, B. Li, J. Hoey, and A. Wong, "Self-supervised pretraining improves performance and inference efficiency in multiple lung ultrasound interpretation tasks," *arXiv preprint arXiv:2309.02596*, 2023.

[24] D. A. Lichtenstein and Y. Menu, "A bedside ultrasound sign ruling out pneumothorax in the critically iii: lung sliding," *Chest*, vol. 108, no. 5, pp. 1345–1348, 1995.

[25] D. A. Lichtenstein, G. Mezière, N. Lascols, P. Biderman, J.-P. Courret, A. Gepner, I. Goldstein, and M. Tenoudji-Cohen, "Ultrasound diagnosis of occult pneumothorax," *Critical care medicine*, vol. 33, no. 6, pp. 1231–1238, 2005.

[26] D. A. Lichtenstein, *Whole body ultrasonography in the critically ill*. Springer Science & Business Media, 2010.

[27] S. Azizi, B. Mustafa, F. Ryan, Z. Beaver, J. Freyberg, J. Deaton, A. Loh, A. Karthikesalingam, S. Kornblith, T. Chen, *et al.*, "Big self-supervised models advance medical image classification," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3478–3488, 2021.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[29] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, *et al.*, "Searching for MobileNetV3," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1314–1324, 2019.

[30] Y. You, J. Li, S. Reddi, J. Hseu, S. Kumar, S. Bhojanapalli, X. Song, J. Demmel, K. Keutzer, and C.-J. Hsieh, "Large batch optimization for deep learning: Training bert in 76 minutes," *arXiv preprint arXiv:1904.00962*, 2019.

A ParenchymalLUS Dataset and Downstream Tasks

ParenchymalLUS is a subset of a larger database of de-identified lung US videos that was partially labelled for previous work [17, 21]. Access to this database was permitted via ethical approval by [redacted].[1] The labelled portion of ParenchymalLUS was split by patient identifier into training, validation, and test sets. Its unlabelled portion consists of 20 000 videos from the unlabelled pool of videos in the database that were predicted to contain a parenchymal view of the lungs by a previously trained lung US view classifier [22]. Below are descriptions of the lung US binary classification tasks for which labels were available in ParenchymalLUS.

A-line vs. B-line Classification (`AB`): A-lines and B-lines are two cardinal artifacts in B-mode lung US that can provide quick information on the status of a patient's lung tissue. A-lines are reverberation artifacts that are indicative of normal, clear lung parenchyma [3]. On lung US, they as horizontal lines deep to the pleural line. Conversely, B-lines are indicative of diseased lung tissue [3]. Generally, the two are mutually exclusive. We evaluate on the binary classification task of A-lines versus B-lines on lung US, as was done in previous work benchmarking joint embedding SSL methods for lung US tasks [23].

Lung Sliding Classification: (`LS`) Lung sliding is a dynamic artifact that, when observed on a parenchymal lung US view, rules out the possibility of a pneumothorax at the site of the probe [24]. The absence of lung sliding is suggestive of pneumothorax, warranting further investigation. On B-mode US, lung sliding manifests as a shimmering of the pleural line [24]. The presence or absence of lung sliding is also appreciable on M-mode lung US images that intersect the pleural line [25, 26]. We evaluate on the binary lung sliding classification task, where positive pairs are 3-second M-mode images originating from the same B-mode video that intersect the pleural line. Following prior studies, we estimate the horizontal bounds of the pleural line using a previously trained object detection model [17] and use the top half of qualifying M-mode images, in decreasing order of total pixel intensity [11].

Table 2 gives the composition of the ParenchymalLUS dataset, along with the number of examples labelled for the `AB` and `LS` tasks.

		UNLABELLED	LABELLED		
			Train	Validation	Test
Total	Patients	5204	1540	330	329
	Videos	20 000	4123	858	936
	Images	4 611 063	927 889	191 437	208 648
`AB` Labels	Videos	–	2100 / 998	441 / 197	512 / 213
	Images	–	484 287 / 216 505	99 132 / 40 608	116 648 / 42 122
`LS` Labels	Videos	–	3169 / 477	631 /103	707 / 96
	Images	–	727 205 / 96 771	146 322 / 23 218	166 753 / 21 911

Table 2: Breakdown of ParenchymalLUS at the video and image level. x / y indicates the number of negative and positive labelled examples available for each task, respectively. Video labels apply to each image within the video. Note that some videos were not labelled for both tasks.

B Sample Weights

For a positive pair of B-mode images occurring at times t_1 and t_2 or M-mode images occurring at positions x_1 and x_2, the sample weight is calculated as follows:

$$w = \frac{\delta_t - |t_2 - t_1| + 1}{\delta_t + 1} \qquad\qquad w = \frac{\delta_x - |x_2 - x_1| + 1}{\delta_x + 1} \tag{1}$$

[1] Omitted to protect anonymity during the review process.

Sample weights were incorporated into each SSL objective trialled in this study. Accordingly, we modified the objective functions for SimCLR, Barlow Twins, and VICReg in order to weigh the contribution to the loss differently based on sample weights. To the authors' knowledge, this study is the first to propose sample weighting schemes for the aforementioned self-supervised learning methods.

The SimCLR objective can be easily modified by multiplying L_i, the per-example NT-Xent loss for the i^{th} positive pair, by sample weight w_i.

$$\mathcal{L}_{\text{SimCLR}} = \frac{1}{N}\sum_{i=1}^{N} w_i L_i \tag{2}$$

For VICReg [20], the invariance term is weighted with w_i for each positive pair in a batch. The invariance term is then calculated as follows:

$$s(Z_1, Z_2) = \frac{1}{N}\sum_{i=1}^{N} w_i \|Z_{1_i} - Z_{2_j}\|_2^2 \tag{3}$$

where Z_1 and Z_2 are batches of predicted embeddings for corresponding positive pairs; that is, Z_{1i} and Z_{2i} correspond to one positive pair. The entire VICReg objective can then be calculated as

$$\mathcal{L}_{\text{VICReg}}(Z_1, Z_2) = \underbrace{\lambda s(Z_1, Z_2)}_{\text{Invariance term}} + \underbrace{\mu(v(Z_1) + v(Z_2))}_{\text{Variance term}} + \underbrace{\nu(c(Z_1) + c(Z_2))}_{\text{Covariance term}} \tag{4}$$

where λ, μ, and ν are weights for each term. Since frames were sampled uniformly at random, $\mathbb{E}[w] \simeq 0.5$. Accordingly, we doubled λ when pretraining VICReg with sample weights.

For the Barlow Twins objective, weighting was applied to each positive pair in the invariance term by computing the weighted normalized cross correlation matrix $\mathcal{C}_W \in \mathbb{R}^{D\times D}$ between the weighted-mean-centered normalized batches of embeddings, Z_1 and Z_2. For a batch of embeddings Z, the calculation for the weighted mean $\bar{Z}$ and standard deviations $\sigma(Z)$ across the batch dimension was performed as follows:

$$\bar{Z} = \frac{\sum_{i=1}^{N} w_i Z_i}{\sum_{i=1}^{N} w_i} \qquad \sigma(Z) = \sqrt{\frac{\sum_{i=1}^{N} w_i (Z_i - \bar{Z})^2}{\sum_{i=1}^{N} w_i}} \tag{5}$$

$$\mathcal{C} = \frac{1}{N}\left(\frac{Z_1 - \bar{Z}_1}{\sigma(Z_1)}\right)^T \left(\frac{Z_2 - \bar{Z}_2}{\sigma(Z_2)}\right) \qquad \sigma(Z) = \sqrt{\frac{\sum_{i=1}^{N} w_i (Z_i - \bar{Z})^2}{\sum_{i=1}^{N} w_i}} \tag{6}$$

where w_i is the sample weight for the i^{th} positive pair in the batch. The redundancy reduction term should still be calculated using the normalized cross correlation matrix $\mathcal{C}$, since its purpose is to decorrelate the embedding dimensions. In the original Barlow Twins, the normalized cross correlation matrix is employed for both terms. The Barlow Twins objective then becomes

$$\mathcal{L}_{\text{BT}} = \underbrace{\sum_{d=1}^{D}(1 - \mathcal{C}_{W_{d,d}})^2}_{\text{Invariance term}} + \underbrace{\lambda \sum_{d=1}^{D}\sum_{\substack{e=1 \\ e\neq d}}^{D} \mathcal{C}_{d,e}{}^2}_{\text{Redundancy reduction term}} \tag{7}$$

C Pretraining and Training Protocols

Unless otherwise stated, all feature extractors are initialized with ImageNet-pretrained weights. Similar studies concentrating on medical imaging have observed that this practice improves downstream performance when compared to random initialization [11, 27]. Moreover, we designate fully supervised classifiers initialized with ImageNet-pretrained weights as a baseline against which to compare models pretrained with SSL.

Evaluation on POCUS follows a similar protocol employed in prior works [8, 10]. Feature extractors with the ResNet18 architecture [28] are pretrained on the Butterfly dataset. Prior to training on the POCUS dataset, a 3-node fully connected layer with softmax activation was appended to the pretrained feature extractor. Five-fold cross validation is conducted with POCUS by fine-tuning the final three layers of the pretrained feature extractor. Unlike prior works, we adopt the average across-folds validation accuracy, instead of taking the accuracy of the combined set of validation set predictions across folds. Presenting the results in this manner revealed the high variance of model performance across folds, which may be due to the benchmark dataset's small video sample size.

All experiments with ParenchymalLUS utilize the MobileNetV3-Small architecture as the feature extractor, which outputs a 576-dimensional representation vector [29]. Feature extractors are pretrained on the union of the unlabelled videos and labelled training set videos in ParenchymalLUS. Performance is assessed via test set classification metrics. Prior to training on the downstream task, a single-node fully connected layer with sigmoid activation was appended to the pretrained feature extractor. We report the performance of linear classifiers trained on the frozen feature extractor's representations, along with classifiers that are fine-tuned end-to-end.

For each joint embedding method, the projectors were multilayer perceptrons with two 768-node layers, outputting 768-dimensional embeddings. Pretraining is conducted for 500 epochs using the LARS optimizer [30] with a batch size of 384 and a learning rate schedule with warmup and cosine decay as in [20].

B-mode and M-mode images were resized to 224×224 and 224×112 pixels respectively using bilinear interpolation. The reason that the width of M-modes was standardized to a smaller value was because the height of B-mode images often far exceeded the number of frames in a 3-second segment of B-mode video. Since feature extractors were initialized with pretrained weights, pixel intensities were mean-centered and normalized using the mean and standard deviation of the ImageNet dataset.

All IVPP pretraining runs were subjected to stochastic data augmentations after intra-video positive pairs were sampled. Each image was subjected to the following sequence of stochastic transformations for data augmentation:

1. Randomly located crop of a fraction of the image's area in the range $[0.4, 1.0]$, followed by resizing to the original image dimensions. For B-mode images, the width/height aspect ratio was confined to the range $[0.8, 1.25]$, while M-mode crops were confined to $[0.4, 0.6]$.
2. With probability 0.5, horizontal reflection.
3. With probability 0.5, random brightness change in the range $[-0.25, 0.25]$.
4. With probability 0.5, random contrast change in the range $[-0.25, 0.25]$.
5. With probability 0.25, random Gaussian blur with a kernel size of 5 pixels and a standard deviation uniformly sampled from the range $[0.1, 2.0]$.

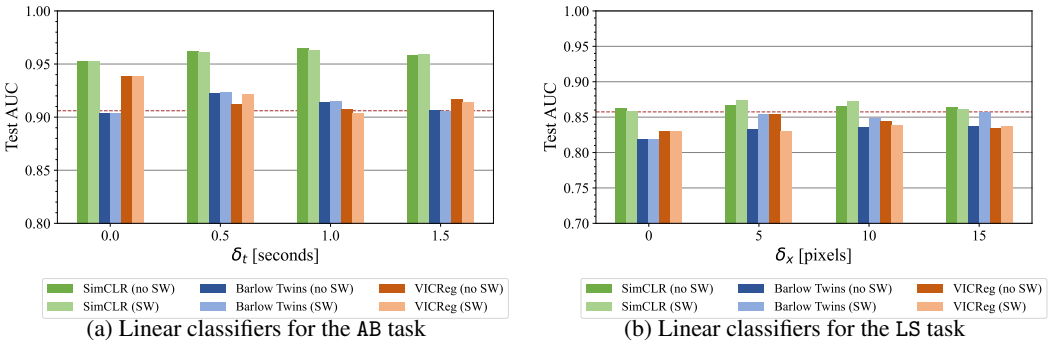
Training runs for the `COVID` and `AB` tasks with B-mode images also utilized this data augmentation pipeline. For `LS` training runs with M-modes, the minimum allowable crop area was increased to 95% of the image's area to ensure the pleural line was almost always visible.

Source code is available at [redacted][2].

[2]Public GitHub repository URL redacted to preserve anonymity.

D Additional Performance Details

Linear Classification Performance: In addition to fine-tuning experiments, linear classifiers were trained using the feature vectors outputted by the pretrained models. Figure 6 summarizes the results obtained for the AB and LS tasks for each of the hyperparameter combinations studied.



(a) Linear classifiers for the AB task

(b) Linear classifiers for the LS task

Figure 6: ParenchymalLUS test set AUC for linear classifiers trained on the AB and LS binary classification tasks and pretrained with a variety of intra-video positive pair thresholds, with and without sample weights (SW). The dashed line indicates initialization with ImageNet-pretrained weights.

Label Efficiency Experiments: Inspection of the central moments and boxplots from each distribution (Figure 5) indicated that the normality and equal variance assumptions for ANOVA were not violated. For each pretraining method, a two-way repeated-measures analysis of variance (ANOVA) was performed to determine whether the mean test AUC scores across values of δ and sample weight usage were different. The independent variables were δ and the presence of sample weights, while the dependent variable was test AUC. Whenever the null hypothesis of the ANOVA was rejected, post-hoc paired *t*-tests were performed to compare the following:

- Pretraining with nonzero δ against standard positive pair selection ($\delta = 0$)
- For the same nonzero δ value, sample weights against no sample weights

For each group of post-hoc tests, the Bonferroni correction was applied to establish a family-wise error rate of $\alpha = 0.05$. To ensure that each training subset was independent, we split the dataset by anonymous patient identifier. This was a necessary step because intra-video images are highly correlated, along with videos from the same patient. As a result, the task became substantially more difficult than naively sampling 5% of training images because the volume *and* heterogeneity of training examples was reduced by training on a small fraction of examples from a small set of patients.

Table 3 gives the mean and standard deviation of each set of trials, for each hyperparameter combination. For each task and each pretraining method, the ANOVA revealed significant interaction effects ($p \leq 0.05$). Accordingly, all intended post-hoc *t*-tests were performed to ascertain (1) which combinations of hyperparameters were different from the baseline setting of augmenting the same frame twice ($\delta = 0$) and (2) values of δ where the addition of sample weights changes the outcome. First, we note that SimCLR was the only pretraining method that consistently outperformed full supervision with ImageNet-pretrained weights. Barlow Twins and VICReg pretraining – both non-contrastive methods – resulted in worse performance.

For the AB task, no combination of intra-video positive pairs or sample weights resulted in statistically significant improvements compared to dual distortion of the same image ($\delta_t = 0$). For Barlow Twins and VICReg, several nonzero δ_t resulted in significantly worse mean test AUC. Sample weights consistently made a difference in Barlow Twins across δ_t values, but only improved mean test AUC for $\delta_t = 1$ and $\delta_t = 1.5$.

Different trends were observed for the LS task. SimCLR with $\delta_x = 5$ and no sample weights improved mean test AUC compared to the baseline where $\delta_x = 0$. No other combination of hyperparameters resulted in a significant improvement. For Barlow Twins, multiple IVPP hyparameter combinations

resulted in improved mean test AUC over the baseline. No IVPP hyperparameter combinations significantly improved the performance of VICReg.

	AB			LS		
Pretrain Method	δ_t	SW	Mean (std) test AUC	δ_x	SW	Mean (std) test AUC
SimCLR	0		0.938 (0.007)	0		0.812 (0.037)
	0.5		0.931 (0.010) *	5		0.824 (0.030) *
	0.5		0.936 (0.007) †	5		0.820 (0.033)
	1		0.934 (0.011)	10		0.815 (0.035)
	1		0.933 (0.011)	10		0.816 (0.037)
	1.5		0.936 (0.013)	15		0.819 (0.034)
	1.5		0.932 (0.012)	15		0.798 (0.039) *†
Barlow Twins	0		0.914 (0.014)	0		0.693 (0.044)
	0.5		0.914 (0.010) *	5		0.694 (0.040)
	0.5		0.883 (0.017) *†	5		0.780 (0.040) *†
	1		0.877 (0.022) *	10		0.705 (0.051)
	1		0.891 (0.018) *†	10		0.706 (0.066)
	1.5		0.870 (0.024) *	15		0.769 (0.037) *
	1.5		0.892 (0.015) *†	15		0.707 (0.071) †
VICReg	0		0.917 (0.011)	0		0.690 (0.042)
	0.5		0.879 (0.024) *	5		0.675 (0.036)
	0.5		0.879 (0.021) *	5		0.679 (0.038)
	1		0.872 (0.023) *	10		0.680 (0.039)
	1		0.876 (0.024) *	10		0.675 (0.040)
	1.5		0.860 (0.026) *	15		0.710 (0.036)
	1.5		0.870 (0.021) *†	15		0.685 (0.039) †
None (ImageNet-pretrained)			0.896 (0.017)			0.783 (0.028)
None (random initialization)			0.774 (0.051)			0.507 (0.022)

* Significantly different ($p < 0.05$) than baseline for the pretraining method where $\delta = 0$

† Significantly different ($p < 0.05$) for particular δ when sample weights are applied, compared to no sample weight

Table 3: ParenchymalLUS test AUC for the the AB and LS tasks when trained using examples from 5% of the patients in the training set. Twenty trials were performed for each pretraining method, value of δ, with and without sample weights (SW). Mean and standard deviation of the test AUC across trials are reported for each condition.