Self-Supervised Ultrasound-Video Segmentation with Feature Prediction and 3D Localised Loss

Edward Ellis[1], Robert Mendel[2], Andrew Bulpitt[1], Nasim Parsa[2], Michael F Byrne[2], and Sharib Ali[1]

[1] School of Computer Science, University of Leeds, Leeds, UK, LS2 9JT
[2] Dova Health Intelligence Inc., Vancouver, Canada, V6B 2W9

Abstract. Acquiring and annotating large datasets in ultrasound imaging is challenging due to low contrast, high noise, and susceptibility to artefacts. This process requires significant time and clinical expertise. Self-supervised learning (SSL) offers a promising solution by leveraging unlabelled data to learn useful representations, enabling improved segmentation performance when annotated data is limited. Recent state-of-the-art developments in SSL for video data include V-JEPA, a framework solely based on feature prediction, avoiding pixel level reconstruction or negative samples. We hypothesise that V-JEPA is well-suited to ultrasound imaging, as it is less sensitive to noisy pixel-level detail while effectively leveraging temporal information. To the best of our knowledge, this is the first study to adopt V-JEPA for ultrasound video data. Similar to other patch-based masking SSL techniques such as VideoMAE, V-JEPA is well-suited to ViT-based models. However, ViTs can underperform on small medical datasets due to lack of inductive biases, limited spatial locality and absence of hierarchical feature learning. To improve locality understanding, we propose a novel 3D localisation auxiliary task to improve locality in ViT representations during V-JEPA pre-training. Our results show V-JEPA with our auxiliary task improves segmentation performance significantly across various frozen encoder configurations, with gains up to 3.4% using 100% and up to 8.35% using only 10% of the training data.

Keywords: Self-Supervised Learning · Ultrasound · Segmentation · ViTs

1 Introduction

Ultrasound (US) imaging is widely used in clinical practice as a low-cost, non-invasive, and portable alternative to CT and MRI. However, building large annotated US datasets is challenging due to high noise, low contrast, and common artefacts like reverberation, acoustic shadowing, and mirror imaging [16]. These factors make US interpretation complex, requiring significant time and expertise, often resulting in high inter-operator variability [3]. US videos support clinical understanding by allowing clinicians to anticipate and identify anatomical structures and pathology across frames, offering contextual information that single-image datasets lack. Video data also aligns more closely with real-world clinical acquisition workflows.

To assist clinician interpretation of US, self-supervised learning (SSL) offers a promising solution [9]. SSL leverages unlabelled data to learn useful representations, improving downstream segmentation with limited labels. SSL has often been applied to US images [8, 7, 10] and recently to US video data [9, 4, 6], harnessing spatial and temporal information. However, many SSL approaches in US imaging present domain-specific pretext learning strategies to improve representation learning, in a contrastive [8, 20] or generative SSL framework [6, 17]. Contrastive learning often requires many negative samples, risking degraded representations from false negatives and demanding large batches or memory banks [13]. Generative SSL on the other hand often emphasises pixel-level reconstruction, increasing susceptibility to noise, with less emphasis on learning high-level structures [13]. The Video Joint Embedding Prediction Architecture (V-JEPA) framework, however, addresses these limitations by avoiding both negative sampling and pixel reconstruction, and instead focussing on abstract representations through masked latent feature prediction. V-JEPA have shown state-of-the-art performance on several natural scene video datasets for classification tasks and demonstrates particular benefit in motion understanding [2].

SSL methods that utilise masked patches, such as VideoMAE [18] or V-JEPA [2], often favour transformer-based models, as positional embeddings provide essential spatial and temporal context for predicting masked regions during pretraining. This is evident in both methods offering pre-trained weights exclusively for Vision Transformer (ViT) models. This poses a challenge for medical imaging uses, with ViTs benefitting from large datasets, performance can suffer in small data scenarios, often performing worse than the convolutional neural network [21]. This drop in performance with ViTs is often attributed to the lack of inductive bias, limited inherent locality and absence of hierarchical feature learning [5]. Several works have proposed techniques to help mitigate these issues. For example, Akkaya *et al.* [1] introduced the LIFE module to incorporate local inductive bias by adding depth-wise separable convolutional layers, which provided local context to ViT embeddings. Liu *et al.* [14] introduced a dense localisation auxiliary task to encourage the ViT to learn spatial relations within an image. In addition architectures like the pyramid vision transformer [19] and SWIN transformer [15] improve hierarchical feature learning through progressive downsampling and local self-attention mechanisms.

We hypothesise that V-JEPA is well suited for ultrasound image segmentation, as its avoidance of pixel-level reconstruction mitigates sensitivity to noise and low contrast problems in US data. Its ability to model spatial and temporal dynamics coherently can help to distinguish between anatomy and artefacts across frames. However, while V-JEPA is well suited to ViT-based models, its performance can suffer in low-data regimes [21]. To address this, we propose a novel 3D localisation auxiliary task that enhances spatial and temporal sensitivity during V-JEPA pretraining, improving ViT's inherent locality limitation in limited data setting. Our approach is model-agnostic, enabling the use of pre-trained weights for domain-specific pretraining without modifying the ViT architectures. Key contributions of our work include:

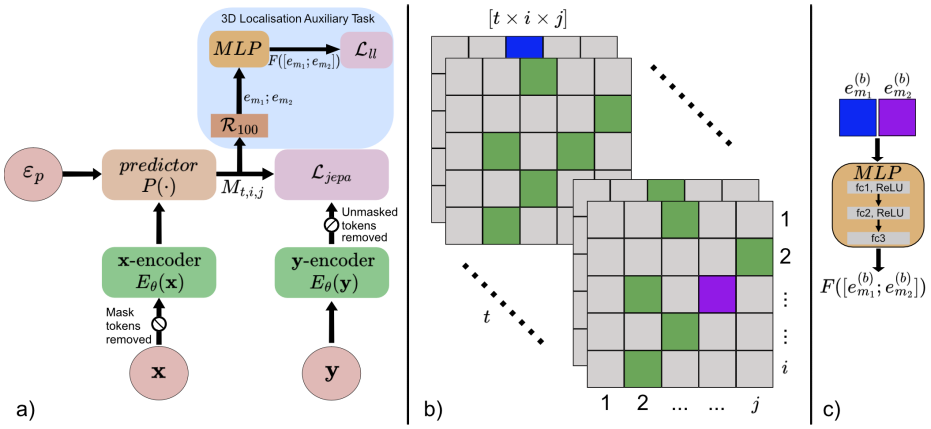

Fig. 1. Block Diagram of our 3D localisation auxiliary task incorporated in the V-JEPA SSL framework. Our auxiliary task takes a random pair of patch embeddings, (e_{m_1}, e_{m_2}), from the predictor and predicts the relative temporal, vertical and horizontal distances between the samples. (a) presents the general SSL framework of V-JEPA with our localisation task. (b) demonstrates the relative localisation between a sample pair (in blue and purple), sampled from the predicted masked areas, $M_{t,i,j}$ (in green). (c) shows the relative localisation prediction, $(F([e^{(b)}_{m_1}; e^{(b)}_{m_2}]))$, from a simple multilayer perceptron (MLP) network between the concatenated sample pair $(e^b_{m_1}; e^b_{m_2})$.

1. Adopting state-of-the-art V-JEPA SSL framework for medical video US image segmentation, in our case cardiac US videos
2. Addressing the inherent locality limitation of ViT-based V-JEPA on small US video data, by improving its locality through a novel learnable 3D localisation auxiliary task.
3. Comprehensive evaluation of video-based SSL techniques on cardiac US videos, including variable dataset sizes.

2 Method

The overall SSL approach proposed is outlined in Figure 1. Below we detail V-JEPA for video segmentation and 3D localisation loss integrated in the V-JEPA framework.

2.1 V-JEPA

V-JEPA builds upon the joint-embedding predictive architecture (JEPA) [12]. JEPA learns by predicting the representations of an input $\mathbf{y}$ from another input $\mathbf{x}$, conditioned on the transformation/corruption between $\mathbf{x}$ and $\mathbf{y}$. In practice this corruption is implemented via masking, which is contextualised for the network through positional embeddings, ϵ_p. We define $\epsilon_p \leftarrow \Delta\mathbf{y}$, where $\Delta\mathbf{y}$ denotes

the spatio-temporal positions of the masked regions of $\mathbf{y}$. The $\mathbf{x}$-encoder, $E_\theta(\mathbf{x})$, is trained on a masked video sequence, outputting an embedding vector for each 'visible' spatio-temporal token. The positional embeddings and output of $\mathbf{x}$-encoder are passed to a predictor network, $P_\phi(\cdot)$, to predict the representation of masked tokens. $P_\phi(\cdot)$ is trained simultaneously to $E_\theta(\mathbf{x})$.

V-JEPA effectively aims to minimise the L_1 loss between predicted and target representations of masked regions (see Eq. 1). Target representations are obtained by using the same encoder on the complete video clip sample, $E_\theta(\mathbf{y})$, with unmasked areas removed. This encoder is not trained (stop gradient sg is used), and updated through the exponential moving average ($\overline{E}_\theta(\cdot)$) of $E_\theta(\mathbf{x})$.

$$\mathcal{L}_{\text{jepa}} = \left\| P_\phi\left(E_\theta(\mathbf{x}), \mathbf{\Delta y}\right) - \text{sg}\left(\overline{E}_\theta(\mathbf{y})\right) \right\|_1 \tag{1}$$

2.2 3D Localisation Auxiliary task

Our localisation task aims to improve spatial understanding during pre-training. We add this task to the outputs of the predictor. Initially, a video clip of T frames and spatial resolution of $H \times W$ is tokenised, each token of shape $2 \times 16 \times 16$. When encoded, this results in a total embedding space of shape $t \times i \times j$ spatio-temporal patch embeddings, $P_{t,i,j}$, where t is the number of tubelets and i and j are the number of spatial tokens. Masked patch embeddings, $M_{t,i,j}$ (predicted patches) are a subset of these tokens, i.e $M_{t,i,j} \subset P_{t,i,j}$. A random subset of 100 concatenated token embedding pairs, denoted $\mathcal{R}_{100}$, is sampled from $M_{t,i,j}$, $\mathcal{R}_{100} \subset M_{t,i,j}$. We compute the 3D normalised target relative translation offset ($\Delta^{(b)}_{m_1,m_2}$) between each randomly sampled embedding pair $e^{(b)}_{m_1}$ and $e^{(b)}_{m_2}$, where m_1 and m_2 correspond to the temporal (t) and spatial (i, j) locations of each embedding. Here b indexes a sample from batch B. This provides a ground truth.

$$\Delta^{(b)}_{m_1,m_2} = \left(\frac{t_1 - t_2}{t},\ \frac{i_1 - i_2}{i},\ \frac{j_1 - j_2}{j} \right),\ i, j, t \in [-1, 1] \tag{2}$$

The sampled embedding pair is concatenated and passed as input to a small MLP, $F(\cdot)$, composed of three fully connected layers. This MLP predicts the relative translation offset, denoted $F([e^{(b)}_{m_1}; e^{(b)}_{m_2}])$. Our local loss, L_{ll}, computes the mean squared error between the predicted and ground truth relative translation offset (see Eq. 3). The overall loss is computed as the average over all pairs in $\mathcal{R}_{100}$, per batch (B), $N_{total} = 100 \times B$. $\mathcal{R}^{(b)}_{100}$ is the set of concatenated sample pairs for sample b.

$$\mathcal{L}_{ll} = \frac{1}{N_{\text{total}}} \sum_{b=1}^{B} \sum_{(m_1,m_2) \in \mathcal{R}^{(b)}_{100}} \left\| F\left(\left[\mathbf{e}^{(b)}_{m_1};\ \mathbf{e}^{(b)}_{m_2} \right] \right) - \Delta^{(b)}_{m_1,m_2} \right\|_2^2 \tag{3}$$

2.3 Combined Loss

Our combined loss is a weighted sum of L_{jepa} (Eq. 1) and L_{ll} (Eq. 3), with λ denoted as weight, shown in Eq. 4. We ablate this λ weighting in Table 1.

$$\mathcal{L}_{\text{combined}} = \lambda \cdot \mathcal{L}_{\text{jepa}} + (1 - \lambda) \cdot \mathcal{L}_{ll} \tag{4}$$

3 Experiments

All experiments were performed on the publicly available CAMUS dataset [11]. CAMUS is a cardiac ultrasound dataset containing clinical exams from 500 patients from University Hospital of St Etienne, France. Data was collected using a GE Vivid E95 ultrasound scanner and GE M5S probe. 2D apical four chamber and two chamber view sequence are provided for each patient containing at least one full cardiac cycle. Annotations for left ventricle endocardium (LV Endo), left ventricle Epicardium (LV Epi) and left atrium wall (LA wall) are provided.

3.1 Experimental Setup

All experiments were implemented using Pytorch and performed on NVIDIA L40S 48GB GPUs. We used a batch-size of 4 during pre-training and downstream training, with 16 frames sampled per video. We used a frame step of 4 and spatial resolution of 224×224 pixels. Published pre-trained weights for both V-JEPA and VideoMAE ViT-L models were used before pre-training on CAMUS dataset. For downstream training after pre-training we freeze the ViT-L encoder, use attentive probing, before passing to a shallow decoder consisting of 2 transpose convolutional layers. This evaluation is similar to described in the V-JEPA paper [2], but with a shallow decoder used to obtain a segmentation output. Pre-training was run for 300 epochs using AdamW with a 20-epoch warmup and a cosine learning rate schedule from 0.0002 to $1e^{-6}$. Downstream training was also run for 300 epochs using AdamW, cross-entropy loss, and a cosine learning rate schedule from $1e^{-3}$ to 0.

3.2 Evaluation Metrics

To evaluate the performance of our method, we used: Dice Similarity Coefficient (DSC = $\frac{2 \cdot |y_{\text{pred}} \cap y_{\text{true}}|}{|y_{\text{pred}}| + |y_{\text{true}}|}$), Jaccard Index (JI = $\frac{|y_{\text{pred}} \cap y_{\text{true}}|}{|y_{\text{pred}} \cup y_{\text{true}}|}$), precision (PPV = $\frac{|y_{\text{pred}} \cap y_{\text{true}}|}{|y_{\text{pred}}|}$) and recall (Rec. = $\frac{|y_{\text{pred}} \cap y_{\text{true}}|}{|y_{\text{true}}|}$). y_{pred} and y_{true} represent the predicted segmentation mask and ground truth segmentation masks, respectively.

4 Results and Discussion

We compare segmentation performance between a supervised and pre-trained ViT-L model. We show results for pre-trained methods: VideoMAE, V-JEPA and V-JEPA with our localisation auxiliary task added. With ViT-L/16 the smallest V-JEPA model available with published weights, we demonstrate the impact when 12 and 16 transformer blocks are frozen during pre-training.

Table 1. Effect of varying λ on validation set. DSC results are shown.

Method	λ Setting			
	0.9	0.75	0.5	0.25
V-Jepa + LL	0.696	0.658	0.708	**0.754**
V-Jepa (12b) + LL	0.782	0.787	0.780	**0.805**
V-Jepa (16b) + LL	0.810	0.812	0.817	**0.818**

4.1 Ablation Results

We include an ablation study to investigate the impact of λ weighting our combined loss function (Eq. 4). Table 1 indicates an optimal λ weighting of 0.25, showing validation set performance using 100% training samples. We found this weighting remains optimal across 10%, 20% and 50% subsets.

4.2 Quantitative Results

Our results in Table 2 show V-JEPA pre-training improves downstream segmentation performance in the CAMUS dataset with all V-JEPA variants outperforming both VideoMAE and the supervised only ViT-L baselines. We show a 10.8% and 14% improvement in DSC by increasing the number of frozen transformer blocks using 100% and 10% training samples respectively. Freezing ViT-L transformer blocks during pre-training reduces over-fitting to the CAMUS dataset by limiting trainable parameters, while leveraging publicly available pre-trained weights for adaptation to our US domain. Furthermore we demonstrate improved performance using our local loss in V-JEPA pre-training. Adding local loss improved DSC using 100% training samples by 1.07% ($p = 1.5e^{-2}$), 3.40% ($p = 2e^{-19}$) and 0.7% ($p = 4.4e^{-3}$) for V-JEPA, V-JEPA (12b) and V-JEPA (16b) configurations respectively. These improvements are all statistically significant with p-values < 0.05. Similarly with 10% training samples we show significant segmentation performance improvement of 7.45% ($p = 2.2e^{-21}$), 8.35% ($p = 2.9e^{-31}$) and 2.31% ($p = 3.2e^{-8}$) for V-JEPA, V-JEPA (12b) and V-JEPA (16b) configurations respectively, with p-values < 0.05. Furthermore Table 2 shows similar improvement in JI results when local loss is added. When inspecting PPV and recall, we see greater improvement to recall with local loss added, particularly as training samples become more limited.

These results demonstrate that adding our local loss auxiliary task benefits V-JEPA pre-training, enhancing representation quality on the small CAMUS dataset, particularly benefitting downstream segmentation performance under limited data scenarios, i.e. with 10% training samples.

4.3 Qualitative Results

Segmentation predictions for 3 example frames are shown for each method in Figure 2 for both 100% and 10% training samples. We can see V-JEPA methods perform better at segmenting each class relative to the ground truth labels. This

Table 2. Quantitative comparison of US video segmentation on CAMUS dataset for different % of training samples. The overall results across all videos in the test set are presented. All models use the ViT-L/16 encoder. SD indicates the standard deviation.

Method	% training samples	DSC ± SD	JI ± SD	PPV ± SD	Recall ± SD
Supervised ViT-L	100	0.681 ± 0.076	0.641 ± 0.116	0.794 ± 0.123	0.746 ± 0.117
	50	0.665 ± 0.080	0.607 ± 0.114	0.766 ± 0.131	0.715 ± 0.119
	20	0.635 ± 0.089	0.571 ± 0.136	0.740 ± 0.148	0.679 ± 0.146
	10	0.605 ± 0.089	0.551 ± 0.138	0.724 ± 0.162	0.653 ± 0.152
VideoMAE	100	0.708 ± 0.076	0.665 ± 0.103	0.797 ± 0.113	0.783 ± 0.112
	50	0.652 ± 0.077	0.615 ± 0.117	0.762 ± 0.133	0.738 ± 0.128
	20	0.643 ± 0.073	0.603 ± 0.119	0.749 ± 0.138	0.725 ± 0.130
	10	0.590 ± 0.080	0.560 ± 0.138	0.714 ± 0.160	0.677 ± 0.155
V-JEPA	100	0.747 ± 0.067	0.679 ± 0.093	0.814 ± 0.100	0.788 ± 0.094
	50	0.729 ± 0.069	0.669 ± 0.097	0.799 ± 0.104	0.788 ± 0.099
	20	0.681 ± 0.088	0.608 ± 0.117	0.753 ± 0.134	0.727 ± 0.123
	10	0.644 ± 0.094	0.574 ± 0.125	0.721 ± 0.147	0.698 ± 0.142
V-Jepa + LL (ours)	100	0.755 ± 0.075	0.674 ± 0.097	0.800 ± 0.107	0.793 ± 0.095
	50	0.747 ± 0.078	0.659 ± 0.097	0.761 ± 0.111	0.814 ± 0.093
	20	0.678 ± 0.081	0.593 ± 0.103	0.757 ± 0.127	0.714 ± 0.115
	10	0.692 ± 0.092	0.600 ± 0.110	0.745 ± 0.130	0.729 ± 0.123
V-JEPA (12b)	100	0.795 ± 0.053	0.736 ± 0.074	0.862 ± 0.077	0.826 ± 0.072
	50	0.775 ± 0.058	0.705 ± 0.077	0.846 ± 0.085	0.797 ± 0.078
	20	0.751 ± 0.035	0.678 ± 0.040	0.810 ± 0.050	0.792 ± 0.053
	10	0.683 ± 0.082	0.609 ± 0.107	0.756 ± 0.125	0.731 ± 0.118
V-Jepa (12b) + LL (ours)	100	0.822 ± 0.048	0.761 ± 0.072	0.874 ± 0.069	0.850 ± 0.068
	50	0.804 ± 0.050	0.743 ± 0.074	0.856 ± 0.076	0.845 ± 0.070
	20	0.774 ± 0.062	0.706 ± 0.089	0.829 ± 0.093	0.819 ± 0.091
	10	0.740 ± 0.069	0.671 ± 0.967	0.812 ± 0.101	0.782 ± 0.105
V-JEPA (16b)	100	0.828 ± 0.044	**0.779 ± 0.064**	0.880 ± 0.064	**0.868 ± 0.067**
	50	0.803 ± 0.051	0.747 ± 0.073	0.861 ± 0.073	0.844 ± 0.074
	20	**0.780 ± 0.068**	**0.707 ± 0.092**	**0.834 ± 0.094**	**0.814 ± 0.089**
	10	0.734 ± 0.079	0.661 ± 0.107	0.812 ± 0.110	0.765 ± 0.117
V-Jepa (16b) + LL (ours)	100	**0.834 ± 0.046**	0.778 ± 0.068	**0.882 ± 0.063**	0.863 ± 0.065
	50	**0.824 ± 0.046**	**0.765 ± 0.068**	**0.870 ± 0.067**	**0.859 ± 0.068**
	20	0.779 ± 0.071	0.704 ± 0.092	0.827 ± 0.098	**0.814 ± 0.094**
	10	**0.751 ± 0.077**	**0.668 ± 0.099**	0.811 ± 0.104	**0.778 ± 0.105**

is demonstrated by smoother class boundaries compared to the supervised ViT and VidMAE methods. The orange double-headed arrow in Figure 2 highlights jagged boundaries in these approaches. Using 100% training samples, adding local loss captures the orientation of the ground truth more effectively compared to each corresponding V-JEPA baseline (highlighted in Figure 2, row 1, with orange brackets). As expected, with much fewer training samples segmentation performance worsens. With 10% training samples, the orientation of the ground truth mask is not captured well across all segmented predictions (see Figure 2, row 4). Secondly, segmented boundaries lose smoothness compared to the ground truth labels. However, the relative size of the overall segmentations are captured best using VJEPA (12b), VJEPA (12b) + LL, VJEPA (16b) and VJEPA (16b) + LL variants, with VJEPA (16b) + LL showing best segmentation results overall, highlighted in orange in Figure 2.

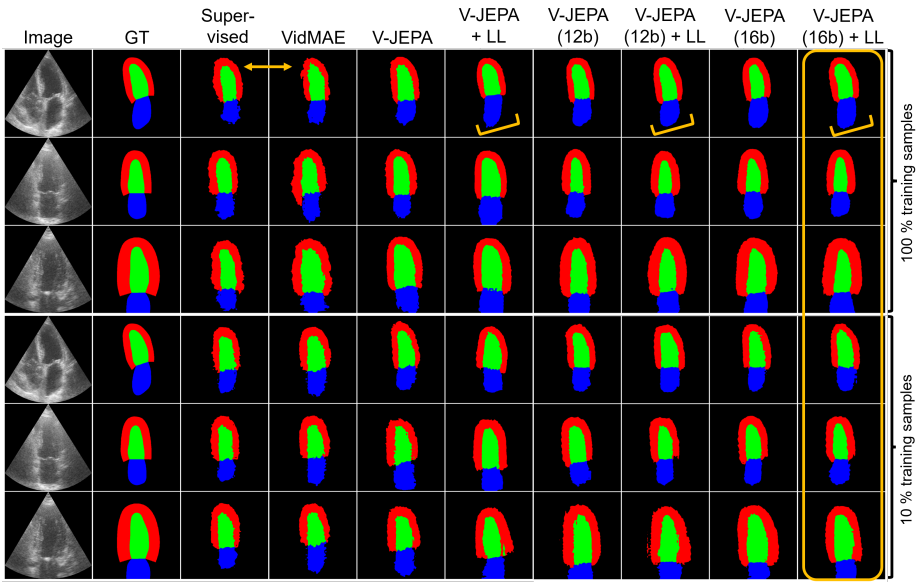

Fig. 2. Qualitative evaluation on CAMUS dataset. 3 example videos were chosen at frame 9. Segmentation predictions across all methods are presented, using 100% and 10% training samples. LV endocardium, LV epicardium, LA wall are indicated in green, red and blue respectively. Orange annotations highlight key points discussed in the qualitative results section (see 4.3).

5 Conclusions

In this work we explored the performance of V-JEPA on cardiac ultrasound data. V-JEPA outperformed the commonly used VideoMAE approach for self-supervised learning on video data. However, with these methods well-suited to transformer-based models, performance can suffer on smaller medical datasets. We proposed a 3D relative localisation auxiliary task to improve V-JEPA pre-training when data is limited. This approach strengthens ViT spatial locality understanding, leading to improved representation learning and significantly better downstream segmentation performance. Future work will apply this method to a broader range of ultrasound video datasets, integrating complementary strategies, such as hierarchical transformers, to enhance performance on small datasets further.

References

1. Akkaya, I.B., Kathiresan, S.S., Arani, E., Zonooz, B.: Enhancing performance of vision transformers on small datasets through local inductive bias incorporation. Pattern Recognition **153**, 110510 (Sep 2024)
2. Bardes, A., Garrido, Q., Ponce, J., Chen, X., Rabbat, M., LeCun, Y., Assran, M., Ballas, N.: Revisiting Feature Prediction for Learning Visual Representations from Video. arXiv (2024), https://arxiv.org/abs/2404.08471
3. Brattain, L.J., Telfer, B.A., Dhyani, M., Grajo, J.R., Samir, A.E.: Machine learning for medical ultrasound: status, methods, and future opportunities. Abdominal Radiology **43**(4), 786–799 (Apr 2018)
4. Chen, L., Rubin, J., Ouyang, J., Balaraju, N., Patil, S., Mehanian, C., Kulhare, S., Millin, R., Gregory, K.W., Gregory, C.R.: Contrastive self-supervised learning for spatio-temporal analysis of lung ultrasound videos. pp. 1–5. IEEE (2023)
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale (2020)
6. E. Lamoureux, S. Ayromlou, S. N. Ahmadi Amiri, H. Rhodin: Segmenting Cardiac Ultrasound Videos Using Self-Supervised Learning. In: 2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). pp. 1–7 (Jul 2023)
7. Ellis, E., Bulpitt, A., Parsa, N., Byrne, M.F., Ali, S.: A Self-Supervised Framework for Improved Generalisability in Ultrasound B-mode Image Segmentation. arXiv preprint arXiv:2502.02489 (2025)
8. Fu, Z., Jiao, J., Yasrab, R., Drukker, L., Papageorghiou, A.T., Noble, J.A.: Anatomy-Aware Contrastive Representation Learning for Fetal Ultrasound. Computer vision - ECCV. European Conference on Computer Vision: proceedings. European Conference on Computer Vision **2022**, 422–436 (Oct 2022)
9. Jiao, J., Droste, R., Drukker, L., Papageorghiou, A.T., Noble, J.A.: Self-Supervised Representation Learning for Ultrasound Video. Proceedings. IEEE International Symposium on Biomedical Imaging **2020**, 1847–1850 (Apr 2020)
10. Jiao, J., Zhou, J., Li, X., Xia, M., Huang, Y., Huang, L., Wang, N., Zhang, X., Zhou, S., Wang, Y.: Usfm: A universal ultrasound foundation model generalized to tasks and organs towards label efficient image analysis. Medical Image Analysis **96**, 103202 (2024)
11. Leclerc, S., Smistad, E., Pedrosa, J., Østvik, A., Cervenansky, F., Espinosa, F., Espeland, T., Berg, E.A.R., Jodoin, P.M., Grenier, T., Lartizien, C., D'hooge, J., Lovstakken, L., Bernard, O.: Deep Learning for Segmentation Using an Open Large-Scale Dataset in 2D Echocardiography. IEEE Transactions on Medical Imaging **38**(9), 2198–2210 (2019)
12. LeCun, Y.: A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. Open Review **62**(1), 1–62 (2022)
13. Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J., Tang, J.: Self-Supervised Learning: Generative or Contrastive. IEEE Transactions on Knowledge and Data Engineering **35**(1), 857–876 (2023)
14. Liu, Y., Sangineto, E., Bi, W., Sebe, N., Lepri, B., Nadai, M.: Efficient training of visual transformers with small datasets. Advances in Neural Information Processing Systems **34**, 23818–23830 (2021)
15. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. pp. 10012–10022 (2021)

16. Quien, M.M., Saric, M.: Ultrasound imaging artifacts: How to recognize them and how to avoid them. Echocardiography **35**(9), 1388–1401 (2018)
17. Szijártó, A., Magyar, B., Szeier, T.A., Tolvaj, M., Fábián, A., Lakatos, B.K., Ladányi, Z., Bagyura, Z., Merkely, B., Kovács, A.: Masked Autoencoders for Medical Ultrasound Videos Using ROI-Aware Masking. pp. 167–176. Springer (2024)
18. Tong, Z., Song, Y., Wang, J., Wang, L.: VideoMAE: masked autoencoders are data-efficient learners for self-supervised video pre-training. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. NIPS '22, Red Hook, NY, USA (2022)
19. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. pp. 568–578 (2021)
20. Zhang, K., Jiao, J., Noble, J.A.: Fetal Ultrasound Video Representation Learning Using Contrastive Rubik's Cube Recovery. In: Simplifying Medical Ultrasound, vol. 15186, pp. 187–197. Springer Nature Switzerland, Cham (2025)
21. Zhu, H., Chen, B., Yang, C.: Understanding why vit trains badly on small datasets: An intuitive perspective. arXiv preprint arXiv:2302.03751 (2023)