带有特征预测和三维局部化损失的自监督超声视频分割

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. 在超声成像中获取和注释大型数据集具有挑战性，因为其对比度低、噪声高且易受伪影影响。这个过程需要大量的时间和临床专业知识。自监督学习（SSL）通过利用未标记的数据学习有用的表示，提供了一种有前途的解决方案，当标注数据有限时，能够提高分割性能。最近在视频数据 SSL 方面的最新发展包括 V-JEPA，这是一个完全基于特征预测的框架，避免了像素级重建或负样本。我们假设 V-JEPA 非常适合超声成像，因为它对噪声像素级细节不太敏感，同时可以有效利用时间信息。据我们所知，这是首次将 V-JEPA 用于超声视频数据的研究。类似于其他基于块的遮掩 SSL 技术，如 VideoMAE，V-JEPA 非常适合基于 ViT 的模型。然而，由于缺乏归纳偏差、有限的空间局部性和层次化特征学习的缺乏，ViT 在小型医学数据集上表现不佳。为了改善对局部性的理解，我们提出了一种新的 3D 定位辅助任务，以在 V-JEPA 预训练期间改善 ViT 表示中的局部性。我们的结果显示，使用我们的辅助任务，V-JEPA 在各种冻结编码器配置中显著改善了分割性能，使用 100% 的训练数据提升高达 3.4%，使用仅 10% 的训练数据提升高达 8.35%。

1 介绍

超声（US）成像在临床实践中广泛应用，作为 CT 和 MRI 的低成本、无创、便携的替代方案。然而，由于高噪声、低对比度以及常见的伪影如混响、声影和镜像成像等因素，构建大型标注的超声数据集具有挑战性。这些因素使得超声解释变得复杂，需要耗费大量时间和专业知识，常导致操作者之间的高度可变性。超声视频通过让临床医生预测并识别跨帧的解剖结构和病理来支持临床理解，提供单张图像数据集所缺乏的上下文信息。视频数据也更符合现实的临床采集工作流程。

为了帮助临床医生解读超声图像，自监督学习（SSL）提供了一种有前途的解决方案 [9] 。SSL 利用无标签数据学习有用的表示，提高在有限标签情况下的下游分割。SSL 经常被应用于超声图像 [8, 7, 10] ，最近也用于超声视频数据 [9, 4, 6] ，利用空间和时间信息。然而，许多在超声成像中的 SSL 方法提出了域特定的前提学习策略来改善表示学习，采用对比 [8, 20] 或生成的 SSL 框架 [6, 17] 。对比学习通常需要许多负样本，面临假负样本导致表征退化的风险，并且需要大批量或记忆库 [13] 。另一方面，生成型 SSL 通常强调像素级重建，增加了噪声的敏感性，并较少关注学习高层结构 [13] 。然而，视频联合嵌入预测架构（V-JEPA）框架通过避免负采样和像素重建解决了这些限制，而是通过掩码的潜在特征预测专注于抽象表示。V-JEPA 在几个自然场景视频数据集上的分类任务中表现出最先进的性能，并在运动理解中展现出特别的优势 [2] 。

www.xueshuxiangzi.com

利用掩码块的 SSL 方法，如 VideoMAE [18] 或 V-JEPA [2] ，通常偏向于基于 transformer 的模型，因为位置嵌入在预训练期间为预测掩码区域提供了关键的空间和时间上下文。这体现在这两种方法都仅为 Vision Transformer (ViT) 模型提供预训练权重。在医学成像应用中，这就构成了一个挑战，因为 ViT 需要大型数据集来体现优势，而在小数据场景中其性能可能受损，往往比卷积神经网络 [21] 更差。在 ViT 中，这种性能下降通常归因于缺乏归纳偏置、有限的内在局部性和缺乏分层特征学习 [5] 。一些研究提出了技术来帮助减轻这些问题。例如，Akkaya 等人 [1] 引入了 LIFE 模块，通过添加深度可分离卷积层来融合局部归纳偏置，为 ViT 嵌入提供局部上下文。Liu 等人 [14] 引入了一项密集定位辅助任务，以鼓励 ViT 学习图像内的空间关系。此外，像金字塔视觉 transformer [19] 和 SWIN transformer [15] 这样的架构通过渐进降采样和局部自注意力机制改善分层特征学习。

我们假设，V-JEPA 非常适合用于超声图像分割，因为它避免像素级重建，从而减轻了 US 数据中对噪声和低对比度问题的敏感性。其连贯建模空间和时间动态的能力有助于区分解剖结构与帧间的伪影。然而，虽然 V-JEPA 适合于基于 ViT 的模型，但其性能可能在数据量少的情形下受损 [21] 。为了解决这个问题，我们提出了一项新颖的 3D 定位辅助任务，在 V-JEPA 预训练过程中增强了空间和时间的敏感性，从而改善了 ViT 在有限数据情况下固有的局部性限制。我们的方法是模型无关的，使得可以在不修改 ViT 架构的情况下，利用预训练权重进行特定领域的预训练。我们的工作贡献包括：

1. 采用最先进的 V-JEPA SSL 框架进行医学视频超声图像分割，在我们的案例中是心脏超声视频
2. 通过一种新颖的可学习 3D 定位辅助任务，提高其局部性，从而解决基于 ViT 的 V-JEPA 在小型超声视频数据上固有的局部性限制。
3. 在心脏超声视频上全面评估基于视频的 SSL 技术，包括不同的数据集规模。

2 方法

提出的整体 SSL 方法概述如图 1 所示。下面我们详细介绍 V-JEPA 视频分割和 V-JEPA 框架中集成的 3D 定位损失。

2.1 V-JEPA

V-JEPA 构建于联合嵌入预测架构 (JEPA) [12] 之上。JEPA 通过在 $\mathbf{x}$ 和 $\mathbf{y}$ 之间的转换/损坏条件下，从另一个输入 $\mathbf{x}$ 预测输入 $\mathbf{y}$ 的表征来学习。在实践中，这种损坏通过掩蔽实现，网络通过位置嵌入 ϵ_p 来对其进行背景化。我们定义 $\epsilon_p \leftarrow \Delta\mathbf{y}$ ，其中 $\Delta\mathbf{y}$ 表示 $\boldsymbol{y}$ 的掩蔽区域的时空位置。$\mathbf{x}$ 编码器 $E_\theta(\mathbf{x})$ 在掩蔽的视频序列上训练，为每个“可见”的时空标记输出嵌入向量。位置嵌入和 $\mathbf{x}$ 编码器的输出传递给预测器网络 $P_\phi(\cdot)$ ，以预测被掩蔽标记的表征。$P_\phi(\cdot)$ 和 $E_\theta(\mathbf{x})$ 同时进行训练。

V-JEPA 主要旨在最小化被遮挡区域的预测表示和目标表示之间的 L_1 损失（参见公式 1 ）。目标表示是通过对完整视频片段样本 $E_\theta(\mathbf{y})$ 使用相同的编码器获得的，其中未被遮挡的区域被去除。此编码器不进行训练（使用停止梯度 sg ），并通过指数移动平均（$\overline{E}_\theta(\cdot)$) 中的 $E_\theta(\mathbf{x})$ ）进行更新。

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

www.xueshuxiangzi.com

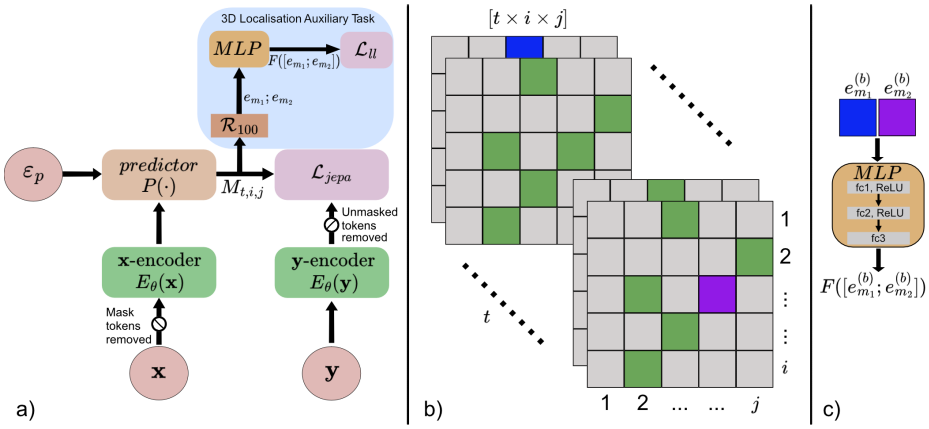

Fig. 1. 我们 3D 定位辅助任务在 V-JEPA SSL 框架中的模块图。我们的辅助任务从预测器中取出一对随机的补丁嵌入，(e_{m_1}, e_{m_2})，并预测样本之间的相对时间、垂直和水平距离。(a) 展示了包含我们定位任务的 V-JEPA 的一般 SSL 框架。(b) 演示了从预测的掩码区域中采样的一对样本（蓝色和紫色）之间的相对定位，$M_{t,i,j}$ （绿色）。(c) 展示了通过一个简单的多层感知器（MLP）网络针对连接的样本对 ($e^b_{m_1};e^b_{m_2}$) 的相对定位预测，($F([e^{(b)}_{m_1};e^{(b)}_{m_2}])$)。

2.2 三维定位辅助任务

我们的定位任务旨在改进在预训练期间的空间理解。我们将此任务添加到预测器的输出中。最初，一个具有 T 帧和空间分辨率为 $H \times W$ 的视频剪辑被标记化，每个标记的形状为 $2 \times 16 \times 16$ 。当编码时，这将产生一个总的嵌入空间，其形状为 $t \times i \times j$ 时空补丁嵌入，$P_{t,i,j}$ ，其中 t 是管块的数量，i 和 j 是空间标记的数量。遮罩补丁嵌入，$M_{t,i,j}$ （预测补丁）是这些标记的一个子集，即 $M_{t,i,j} \subset P_{t,i,j}$ 。从 $M_{t,i,j}$ ，$\mathcal{R}_{100} \subset M_{t,i,j}$ 中随机采样 100 个拼接的标记嵌入对，记为 $\mathcal{R}_{100}$ 。我们计算每对随机采样嵌入 $e^{(b)}_{m_1}$ 和 $e^{(b)}_{m_2}$ 之间的 3D 归一化目标相对平移偏差（$\Delta^{(b)}_{m_1,m_2}$ ），其中 m_1 和 m_2 分别对应每个嵌入的时间（t ）和空间（i,j ）位置。这里 b 索引批次 B 中的一个样本。这提供了一个真实值。

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

采样的嵌入对被连接后作为输入传递给一个小型 MLP，$F(\cdot)$ ，由三个全连接层组成。这个 MLP 预测相对的平移偏移，记作 $F([e^{(b)}_{m_1};e^{(b)}_{m_2}])$ 。我们的局部损失，L_{ll} ，计算预测的和平移偏移的真实值之间的均方误差（见公式 3 ）。总体损失计算为 $\mathcal{R}_{100}$ 中所有对的批次（B ）的平均值，$N_{total} = 100 \times B$ 。$\mathcal{R}^{(b)}_{100}$ 是样本 b 的连接样本对的集合。

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

www.xueshuxiangzi.com

Table 1. 在验证集上变化 λ 的效果。显示了 DSC 结果。

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

我们的组合损失是 L_{jepa} （方程 1 ）和 L_{ll} （方程 3 ）的加权和，其中 λ 表示权重，如方程 4 所示。我们在表 1 中对这个 λ 加权进行了消融分析。

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

所有实验都是在公开的 CAMUS 数据集上进行的。CAMUS 是一个心脏超声数据集，包含来自法国圣艾蒂安大学医院的 500 名患者的临床检查。数据是使用 GE Vivid E95 超声扫描仪和 GE M5S 探头收集的。对于每位患者，提供了 2D 心尖四腔和二腔视图序列，包含至少一个完整的心动周期。提供了左心室内膜（LV Endo）、左心室外膜（LV Epi）和左心房壁（LA wall）的标注。

所有实验均使用 Pytorch 实现，并在 NVIDIA L40S 48GB GPUs 上进行。我们在预训练和下游训练期间使用了 4 的批量大小，每个视频采样 16 帧。我们使用帧步为 4 ，空间分辨率为 224×224 像素。在对 CAMUS 数据集进行预训练之前，使用了 V-JEPA 和 VideoMAE ViT-L 模型的已发布预训练权重。在预训练后的下游训练中，我们冻结 ViT-L 编码器，使用注意探测，然后传递给由 2 个转置卷积层组成的浅解码器。此评估类似于 V-JEPA 论文中描述的 [2] ，但使用浅解码器以获得分割输出。预训练运行了 300 个时期，使用 AdamW，采用了 20 个时期的预热以及从 0.0002 到 $1e^{-6}$ 的余弦学习率调度。下游训练也运行了 300 个时期，使用 AdamW、交叉熵损失以及从 $1e^{-3}$ 到 0 的余弦学习率调度。

为了评估我们方法的性能，我们使用了：Dice 相似系数（DSC $= \frac{2 \cdot |y_{\text{pred}} \cap y_{\text{true}}|}{|y_{\text{pred}}| + |y_{\text{true}}|}$ ）、Jaccard 指数（JI $= \frac{|y_{\text{pred}} \cap y_{\text{true}}|}{|y_{\text{pred}} \cup y_{\text{true}}|}$ ）、精确率（PPV $= \frac{|y_{\text{pred}} \cap y_{\text{true}}|}{|y_{\text{pred}}|}$ ）和召回率（Rec. $= \frac{|y_{\text{pred}} \cap y_{\text{true}}|}{|y_{\text{true}}|}$ ）。y_{pred} 和 y_{true} 分别代表预测的分割掩膜和真实标签的分割掩膜。

我们比较了一个监督和预训练的 ViT-L 模型之间的分割性能。我们展示了预训练方法的结果：VideoMAE、V-JEPA 以及添加了我们定位辅助任务的 V-JEPA。在使用 ViT-L/16 作为最小的 V-JEPA 模型（该模型的权重已发布）的情况下，我们展示了预训练期间冻结 12 个和 16 个 transformer 模块时的影响。

我们进行了一项消融研究，以调查 λ 对我们组合损失函数（公式 4 ）加权的影响。表 1 指出了 0.25 的最佳 λ 加权，展示了使用 100 % 训练样本的验证集性能。我们发现此加权在 10 %、20 % 和 50 % 子集上仍然是最优的。

我们的结果在表 **??** 中显示，V-JEPA 预训练提升了在 CAMUS 数据集上的下游分割性能，所有 V-JEPA 变体都优于 VideoMAE 和仅监督的 ViT-L 基线。当使用 100 % 和 10 % 训练样本时，通过增加冻结的变压器块数量，我们分别在 DSC 上提高了 10.8 % 和 14 %。预训练期间冻结 ViT-L 变压器块，通过限制可训练参数减少对 CAMUS 数据集的过拟合，同时利用公开可用的预训练权重适

www.xueshuxiangzi.com

应我们的 US 领域。此外，我们展示了在 V-JEPA 预训练中使用我们的局部损失提高了性能。添加局部损失使得在使用 100 % 训练样本时，V-JEPA、V-JEPA (12b) 和 V-JEPA (16b) 配置的 DSC 分别提高了 1.07 % ($p = 1.5e^{-2}$)、3.40 % ($p = 2e^{-19}$) 和 0.7 % ($p = 4.4e^{-3}$)。这些改进在统计上都是显著的，p -值皆小于 0.05。同样地，使用 10 % 训练样本，我们展示了显著的分割性能提升：V-JEPA、V-JEPA (12b) 和 V-JEPA (16b) 配置分别为 7.45 % ($p = 2.2e^{-21}$)、8.35 % ($p = 2.9e^{-31}$) 和 2.31 % ($p = 3.2e^{-8}$)，且 p -值皆小于 0.05。此外，表 **??** 显示，当加入局部损失时，JI 结果有类似的改进。在检查 PPV 和召回时，我们看到，添加局部损失后，召回得到更大改善，特别是在训练样本更有限的时候。

这些结果表明，添加我们的局部损失辅助任务有助于 V-JEPA 预训练，尤其在小型 CAMUS 数据集上提高表示质量，特别是在有限数据场景下改善下游分割性能，即仅使用 10% 训练样本时。

对于每种方法，3 个示例帧的分割预测如图 2 所示，其中分别使用 100 % 和 10 % 的训练样本。我们可以看到，相对于真实标签，V-JEPA 方法在分割每个类别时表现更好。这通过与监督的 ViT 和 VidMAE 方法相比更平滑的类别边界得到证明。图 2 中的橙色双头箭头突出了这些方法中的锯齿边界。使用 100 % 训练样本，添加局部损失更有效地捕捉了真实标签的方向性，相较于每个对应的 V-JEPA 基线（在图 2 的第 1 行用橙色括号突出显示）。如预期，在训练样本大幅减少时，分割性能变差。使用 10 % 训练样本时，真实标签掩膜的方向在所有分割预测中都未被很好地捕捉（见图 2 ，第 4 行）。其次，与真实标签相比，分割边界失去了平滑性。然而，整体分割的相对尺寸在使用 VJEPA (12b)、VJEPA (12b) + LL、VJEPA (16b) 和 VJEPA (16b) + LL 变体时捕捉得最好，其中 VJEPA (16b) + LL 显示出总体上最佳的分割结果，在图 2 中以橙色突出显示。

在这项工作中，我们探讨了 V-JEPA 在心脏超声数据上的表现。V-JEPA 在视频数据自监督学习方面优于常用的 VideoMAE 方法。然而，由于这些方法适合于基于 transformer 的模型，其在较小的医学数据集上的性能可能会受到影响。我们提出了一项 3D 相对定位辅助任务，以改善数据有限情况下 V-JEPA 的预训练。这一方法增强了 ViT 的空间定位理解能力，从而改善了表示学习，并显著提高了后续分割性能。未来的工作将把这一方法应用于更广泛的超声视频数据集，并整合诸如分层 transformer 等补充策略，以进一步增强小数据集上的性能。

www.xueshuxiangzi.com

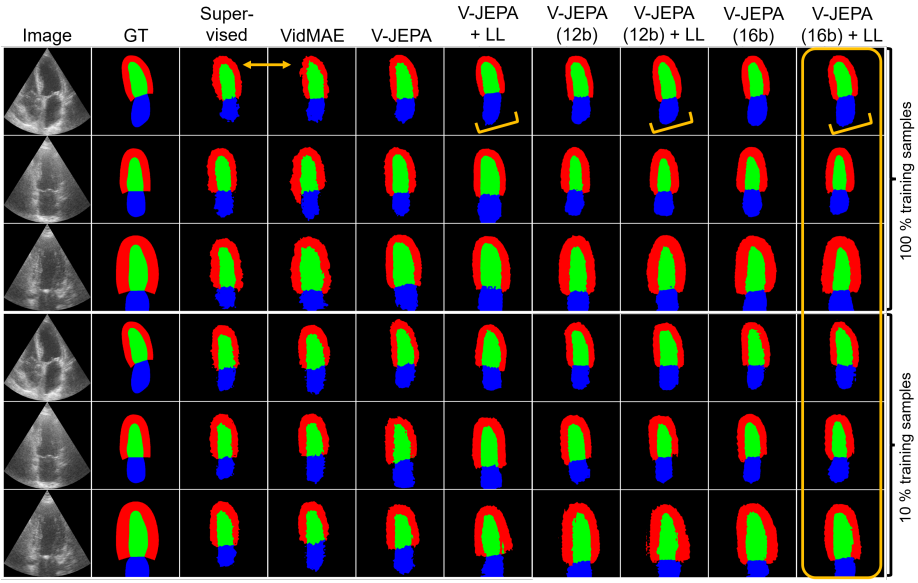

Fig. 2. 对 CAMUS 数据集的定性评估。在第 9 帧选择了 3 个示例视频。展示了使用 100 个% 和 10 个% 训练样本的所有方法的分割预测。左心室心内膜、左心室心外膜、左心房壁分别用绿色、红色和蓝色表示。橙色注释突出了定性结果部分中讨论的关键点（见 **??** ）。

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)

www.xueshuxiangzi.com

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)

www.xueshuxiangzi.com