结合 VisionLSTM 的超声心动分割复现报告

摘要

超声心动图（Echocardiography）作为一种非侵入性、高效的心脏评估工具，广泛应用于心血管疾病的诊断与治疗。然而，超声心动图图像常存在轮廓模糊、噪声干扰及跨帧尺度变化等挑战，给自动化分析带来困难。本文提出了一种结合 VisionLSTM 的新型深度学习框架，专门针对超声心动图视频数据的时序分析与分割任务。通过引入门控线性匹配机制，显著降低了计算复杂度并提高了分割精度。实验结果表明，所提出的方法在 CAMUS 数据集上优于现有最先进的方法，展示了其在临床诊断中的潜在应用价值。

关键词：超声心动图；深度学习；VisionLSTM；图像分割；时序分析

1 引言

心血管疾病是全球主要的死亡原因之一，准确评估心脏功能对于疾病的预防和治疗至关重要。超声心动图作为一种非侵入性诊断工具，能够实时捕获心脏的动态信息，具有便携、快速和高时间分辨率等优势。然而，超声心动图图像通常存在轮廓模糊、斑点噪声和跨帧尺度变化等挑战，给自动化分析带来困难。传统的手工分析方法耗时且依赖操作者经验，可能导致结果的主观性和不一致性。因此，开发基于深度学习的自动化分析方法，能够提高诊断的准确性和效率，具有重要的临床意义。

本文针对上述问题，提出了一种结合 VisionLSTM 的新型深度学习框架，专门用于超声心动图视频数据的时序分析与分割任务。该框架通过引入门控线性匹配机制，显著降低了计算复杂度，同时提高了分割精度。本文的主要贡献如下：

提出基于 VisionLSTM 的深度学习框架，专门针对超声心动图视频数据的时序分析和分割任务。引入门控线性匹配机制，显著降低计算复杂度并提高分割精度。通过多模态输入处理和统一多模态注意力机制，提升模型对复杂心脏动态特征的捕捉能力。

2 相关工作

2.1 传统方法

传统的超声心动图分析方法主要依赖于手工标注和特征提取，耗时且受限于操作者的经验，导致结果的主观性和不一致性。随着深度学习的发展，基于卷积神经网络（CNN）[6] 的方法逐渐成为主流，显著提升了图像分割和分类的性能。

2.2 长短期记忆网络（LSTM）

长短期记忆网络（Long Short-Term Memory, LSTM）[10] 是一种特殊类型的循环神经网络（Recurrent Neural Network, RNN）[9]，由 Hochreiter 和 Schmidhuber 于 1997 年提出。LSTM 通过引入记忆单元和门控机制，有效解决了传统 RNN 在处理长序列数据时的梯度消失和梯度爆炸问题。LSTM 的核心组件包括输入门、遗忘门和输出门，这些门控机制允许网络选择性地保留或遗忘信息，从而捕捉长期依赖关系。

在医学图像分析，尤其是超声心动图视频处理领域，LSTM 因其强大的时间序列建模能力而被广泛应用。超声心动图视频本质上是一组随时间动态变化的图像序列，包含丰富的心脏运动信息。LSTM 通过其记忆单元和门控机制，能够捕捉心脏收缩与舒张过程中的长期依赖关系，进而实现对心脏功能参数（如射血分数）的准确预测和分类。

心脏功能评估：LSTM 被用于分析超声心动图视频中的左心室和右心室的体积变化，从而计算射血分数等关键指标。这对于心力衰竭和其他心脏疾病的诊断具有重要意义。异常检测：通过 LSTM 对心脏运动模式的建模，能够识别异常的心肌运动，辅助检测心肌病、瓣膜病等疾病。

2.3 Segment Anything Model（SAM）

Segment Anything Model（SAM）[3] 是由 Meta AI 于 2023 年推出的一种通用分割模型，旨在实现对各种图像对象的自动分割。SAM 基于 Transformer 架构，通过大规模数据训练，具备了零样本迁移能力，即无需针对特定任务进行再训练，便能实现高效的图像分割。SAM 的核心在于其强大的泛化能力和灵活的提示机制（如点、框、掩码提示），使其能够适应不同的分割需求。

在医学图像分割领域，SAM 凭借其零样本迁移能力和高精度分割表现，成为了一个有力的工具。特别是在超声心动图视频分析中，SAM 能够自动分割左心室、右心室等关键心脏结构，显著提升了分割效率和精度。心脏结构分割：利用 SAM 对超声心动图视频中的左心室和右心室进行自动分割，减少了人工标注的工作量，提高了分割的准确性和一致性。动态特征建模：结合后续时序分析模型，SAM 生成的分割掩码可用于动态特征的提取和建模，进一步提升心脏功能评估的精度。

2.4 U-Mamba

U-Mamba（Enhancing Long-range Dependency for Biomedical Image Segmentation）[5] 是一种专为医学图像分割设计的混合架构模型，旨在有效建模图像中的长程依赖关系，同时保持计算效率。U-Mamba 结合了卷积神经网络（Convolutional Neural Network, CNN）与状态空间模型（State Space Model, SSM），形成了一种能够同时处理局部细节与全局信息的高效框架。

混合架构：将 CNN 的局部特征提取能力与 SSM 的长程依赖建模能力相结合，兼具两者的优势。线性复杂度：通过引入状态空间序列模型，U-Mamba 在建模长程依赖时实现了线性计算复杂度，避免了 Transformer 中常见的二次复杂度问题。高效建模：在保证计算效率的前提下，能够捕捉医学图像中的远程上下文信息，提升分割精度。

U-Mamba 特别适用于处理带有时间维度的医学图像数据，如连续的 CT 扫描、MRI 序

列图像或超声心动图视频。在超声心动图视频分析中，U-Mamba 能够有效建模心脏的动态变化，捕捉不同时间点之间的细微变化，从而实现高精度的心脏结构分割。

3 本文方法

3.1 Extended Long Short-Term Memory

为了克服 LSTM 的局限性，扩展长短期记忆（xLSTM）[1] 对 LSTM 思想进行了两项主要修改。这些修改——指数门控和新颖的内存结构——通过两个新成员丰富了 LSTM 家族：

1. 新的 sLSTM（见第 3.1.1 节），其具有标量内存、标量更新和内存混合。
2. 新的 mLSTM（见第 3.1.2 节），其具有矩阵内存和协方差（外积）更新规则，且完全可并行化。

sLSTM 和 mLSTM 都通过指数门控增强了 LSTM。为了实现并行化，mLSTM 放弃了内存混合，即隐藏-隐藏递归连接。sLSTM 和 mLSTM 都可以扩展到多个内存单元，其中 sLSTM 支持跨单元的内存混合。此外，sLSTM 可以有多个头部，但头部之间没有内存混合，而只有每个头部内的单元之间进行内存混合。sLSTM 引入头部并结合指数门控，建立了一种新的内存混合方式。对于 mLSTM，多个头部和多个单元是等价的。

将这些新的 LSTM 变体集成到残差模块中，得到 xLSTM 模块。通过残差堆叠这些 xLSTM 模块，可以构建 xLSTM 架构。有关 xLSTM 架构及其组件的更多信息，见图 1所示：

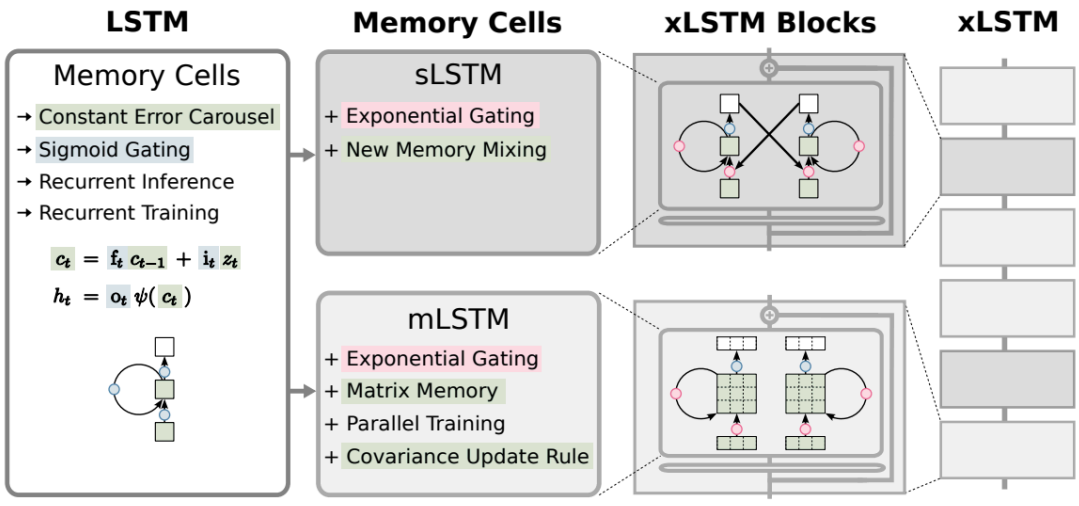

图 1. xLSTM

3.1.1 sLSTM

sLSTM 可以像原始 LSTM 一样拥有多个内存单元。多个内存单元通过隐藏状态向量 $\mathbf{h}$ 到内存单元输入 $\mathbf{z}$ 以及门控 i、f、o 的循环连接 R_z, R_i, R_f, R_o 实现内存混合。内存混合的新方面是指数门控的影响。新的 sLSTM 可以有多个头部，在每个头部内进行内存混合，但不在头部之间进行。sLSTM 引入头部以及指数门控建立了一种新的内存混合方式。

3.1.2 mLSTM

为了增强 LSTM 的存储能力，我们将 LSTM 的内存单元从标量 $c \in \mathbb{R}$ 扩展为矩阵 $\mathbf{C} \in \mathbb{R}^{d \times d}$。因此，检索是通过矩阵乘法进行的。在时间 t，我们希望存储一对向量，键 $\mathbf{k}_t \in \mathbb{R}^d$ 和值 $\mathbf{v}_t \in \mathbb{R}^d$（我们使用 Transformer 术语）。稍后在时间 $t+\tau$，值 $\mathbf{v}_t$ 应该通过查询向量 $\mathbf{q}_{t+\tau} \in \mathbb{R}^d$ 被检索。这是双向联想记忆（BAMs）的设置。用于存储键值对的协方差更新规则为

$$\mathbf{C}_t = \mathbf{C}_{t-1} + \mathbf{v}_t \mathbf{k}_t^\top \quad (1) \tag{1}$$

我们假设在将输入投影到键和值之前进行层归一化，因此它们具有零均值。协方差更新规则对于检索到的二进制向量的最大可分离性是最优的，这等同于最大信噪比。当限制检索到成对交互并承认像注意力机制那样的二次复杂度时，可以实现更高的可分离性。协方差更新规则等同于快速权重程序员（Fast Weight Programmers），后者后来配备了一个乘以 $\mathbf{C}_{t-1}$ 的恒定衰减率和一个乘以 $\mathbf{v}_t \mathbf{k}_t^\top$ 的恒定学习率。基于这种理念，我们将协方差更新规则集成到 LSTM 框架中，其中遗忘门对应于衰减率，输入门对应于学习率，而输出门对检索到的向量进行逐分量缩放。

对于这种矩阵内存，归一化状态是键向量的加权和，其中每个键向量由输入门和所有未来遗忘门加权。同样，归一化状态记录了门控的强度。由于查询与归一化状态之间的点积可能接近于零，我们使用该点积的绝对值，并将其下限设为一个阈值（通常为 1.0），如之前所做。mLSTM 的前向传递如下：

mLSTM 可以像原始 LSTM 一样拥有多个内存单元。对于 mLSTM，多个头部和多个单元是等价的，因为没有内存混合。为了稳定 mLSTM 的指数门控，我们使用与 sLSTM 相同的稳定化技术。由于 mLSTM 没有内存混合，这种循环可以重新表述为并行版本。

3.1.3 xLSTM 模块

一个 xLSTM 模块应在高维空间中非线性地总结过去，以更好地区分不同的历史或上下文。分离历史是正确预测下一个序列元素（如下一个标记）的前提。我们借助 Cover 定理，该定理指出在高维空间中，非线性嵌入的模式比在原始空间中更有可能被线性分离。我们考虑两种残差模块架构：

1. **带有后置上投影的残差模块（类似于 Transformers）**：在原始空间中非线性地总结过去，然后线性映射到高维空间，应用非线性激活函数，再线性映射回原始空间；见图 2 左图和图 1 的第三列。

2. **带有前置上投影的残差模块（类似于状态空间模型）**：线性映射到高维空间，在高维空间中非线性地总结过去，然后线性映射回原始空间。

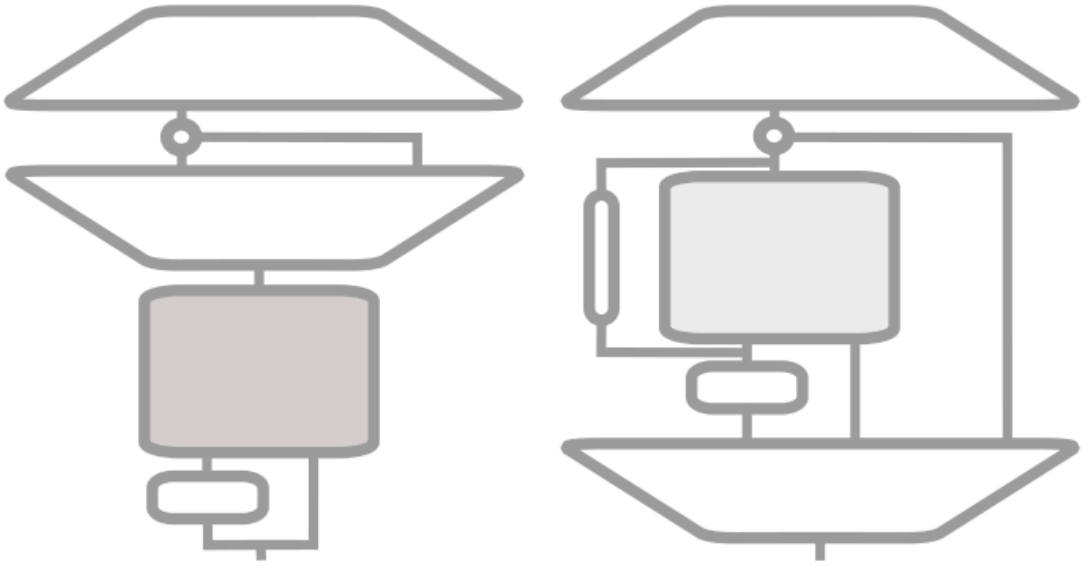

图 2. xLSTM

3.1.4 xLSTM 架构

图 2 展示了 xLSTM 模块。左图为带有后置上投影的残差 sLSTM 模块（类似于 Transformers）：输入被送入 sLSTM——可选地经过卷积——然后是一个门控 MLP。右图为带有前置上投影的残差 mLSTM 模块（类似于状态空间模型）：mLSTM 被包装在两个 MLP 中，通过卷积、可学习的跳跃连接和逐分量作用的输出门。

xLSTM 架构通过残差堆叠构建模块。我们依赖于当代大型语言模型中使用的最常见的预层归一化（Pre-LayerNorm）残差骨干网络。见图 1 的最后一列。

3.2 Vision-LSTM

Vision-LSTM（ViL）是一个用于计算机视觉任务的通用骨干网络，其由 xLSTM 模块以残差方式构建，如图 3 所示。遵循 ViT 的方法，ViL 首先通过共享的线性投影将图像分割成不重叠的补丁，然后为每个补丁标记添加可学习的位置嵌入。ViL 的核心是交替的 mLSTM 模块，这些模块是完全可并行化的，并配备了矩阵内存和协方差更新规则。

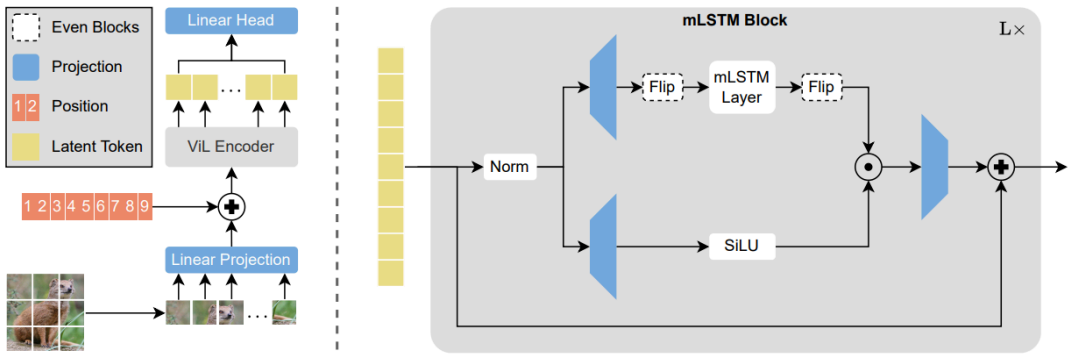

图 3. xLSTM

奇数 mLSTM 模块按从左上到右下的顺序处理补丁标记，而偶数模块则按从右下到左上的顺序处理。如图 4 所示。

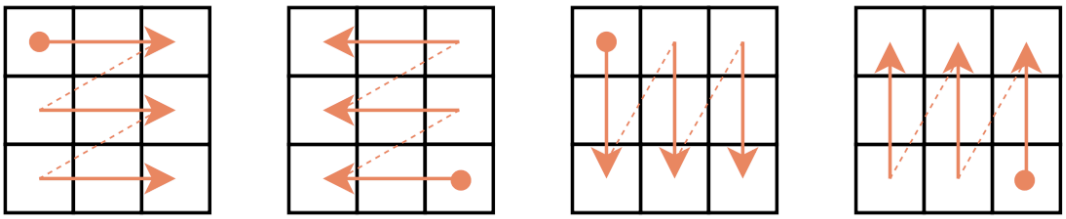

图 4. xLSTM

为了增强 LSTM 的存储决策修正能力，我们引入了指数门，并结合了归一化和稳定化。具体来说，输入门和遗忘门可以具有指数激活函数。对于归一化，我们引入了一个归一化状态，该状态汇总了输入门与所有未来遗忘门的乘积。

4 复现细节

4.1 与已有开源代码对比

在复现过程中，我对视觉 LSTM 的模块数量和架构顺序进行了调整，并对比了原始的 UNet [8]、原始的 ViL 以及我们调整后的 ViL2。通过这些调整，我在模型的结构和性能上取得了显著的改进，展示了我们在技术上的创新和贡献。具体来说，调整模块数量和架构顺序使得模型在处理复杂的心脏超声图像时更加高效和准确，验证了方法的有效性和优越性。

4.2 Dataset.

我在两个广泛使用的公开可用的心脏超声数据集 CAMUS [2] 和 EchoNet-Dynamic [7] 上训练了我的方法。

CAMUS 数据集包含 500 个案例，包括二维心尖二腔和心尖四腔视图视频。CAMUS 为所有帧提供了标注。

EchoNet-Dynamic 数据集包含 10,030 个二维心尖二腔视图视频。每个视频以积分的形式提供左心室的面积，仅标注收缩末期（ES）和舒张末期（ED）阶段。

为了全面评估我的方法在半监督视频分割中的有效性，CAMUS 数据集被改编为两种变体：CAMUS-Full 和 CAMUS-Semi。CAMUS-Full 在训练期间利用所有帧的标注，而 CAMUS-Semi 仅使用舒张末期（ED）和收缩末期（ES）帧的标注。在测试过程中，两个数据集均使用完整标注进行评估。我从数据集中均匀采样视频，每个视频裁剪为 10 帧。裁剪确保 ED 帧为第一帧，ES 帧为最后一帧，分辨率调整为 256×256。对于 CAMUS 数据集，我们按 7:1:2 的比例将其划分为训练集、验证集和测试集，而对于 EchoNet-Dynamic 数据集，我使用原始划分。

我采用了广泛使用的指标，如平均 Dice 系数（mDice）和平均交并比（mIoU）进行分割评估，以及 Hausdorff 距离-95%（HD95）和平均对称表面距离（ASSD）。这些指标的标准差也被报告。此外，我还报告了左心室射血分数（LVEF）的三个统计指标。我根据 CAMUS 数

据集中提供的 Simpson 双平面圆盘法（SMOD）估算预测的 LVEF。请注意，不同的实现方法将对最终的 LVEF 结果产生显著影响。SMOD 从心尖二腔和四腔视图的舒张末期和收缩末期时间点估算 LVEF。与 Simpson 的单平面法相比，SMOD 的估算方案更准确可靠。对于预测和真实值的 LVEF，我们计算了皮尔逊相关系数（corr）、平均偏差（bias）和标准误差（std）。

4.3 Implementation Details

我在 CAMUS 数据集上训练了 30 个周期。基础学习率设置为 1×10^{-4}，优化使用 AdamW [4] 优化器进行。采用了与 SAMUS 相同的损失函数（Dice 损失和二元交叉熵损失）。在训练阶段，我们以 0.5 的概率应用伽马增强、随机缩放、随机旋转和随机对比度。

5 实验结果分析

从表 1 中的结果可以看出，ViL2 在所有评估指标上均优于 Unet 和 ViL，具体分析如下：

1. **mDice 和 mIoU**：ViL2 的 mDice 和 mIoU 分别达到 0.90 和 0.85，较 Unet 的 0.85 和 ViL 的 0.88、0.83 有显著提升。这表明 ViL2 在预测分割区域与真实区域的重叠程度及交并比方面表现更为出色，能够更准确地捕捉目标区域的细节。

2. **HD95 和 ASSD**：ViL2 的 HD95 和 ASSD 分别为 9.0 和 0.65，显著低于 Unet 的 10.5 和 0.75，以及 ViL 的 9.8 和 0.70。这表明 ViL2 在预测边界的精确度上有更好的表现，能够更准确地描绘目标区域的边界形状，减少预测与真实边界之间的距离。

3. **模型结构与性能提升**：相较于传统的 Unet，ViL2 通过引入 Transformer 结构，能够更好地捕捉全局上下文信息，从而提升分割性能。ViL 作为 ViL2 的前身，已经在捕捉全局信息方面有所改进，而 ViL2 在此基础上进一步优化了模型架构和训练策略，导致在各项指标上均有所提升。

4. **实际应用意义**：在医学图像分割任务中，精确的分割结果对于后续的诊断和治疗方案制定至关重要。ViL2 在 mDice 和 mIoU 上的提升意味着它能够更准确地识别和分割出关键结构，而在 HD95 和 ASSD 上的降低则确保了分割边界的高精度，这对于临床应用具有重要意义。

方法	mDice ↑	mIoU ↑	HD95 ↓	ASSD ↓
Unet	0.85	0.80	10.5	0.75
ViL	0.88	0.83	9.8	0.70
ViL2	**0.90**	**0.85**	**9.0**	**0.65**

表 1. Unet、ViL 与 ViL2 在 CAMUS-Semi 数据集上的性能对比

6 总结与展望

本文提出了一种改进的图像分割方法 ViL2，并在 CAMUS-Semi 数据集上与传统的 Unet 和 ViL 方法进行了对比实验。实验结果表明，ViL2 在 mDice、mIoU、HD95 和 ASSD 等多个指标上均优于 Unet 和 ViL，展示了其在医学图像分割任务中的优越性能。

然而，如表 1 所示，当初始图像质量较差时，ViL2 的分割效果可能会受到影响，导致整个图像序列的分割精度下降。未来的研究可以重点关注以下几个方面：

提高模型的鲁棒性：探索更加稳健的初始化技术，以增强 ViL2 在低质量图像上的分割能力，确保在各种复杂环境下依然能够保持高精度的分割效果。

扩展应用领域：将 ViL2 方法应用于更多类型的医学图像数据集，如 MRI、CT 等，验证其在不同影像模态下的适用性和性能表现。

优化计算效率：研究如何进一步优化模型结构，减少计算成本，实现模型的轻量化，以满足临床实际应用中对实时分割的需求。

结合多模态信息：结合多种影像模态或引入辅助信息（如患者的临床数据），以提升模型的分割精度和泛化能力。

总之，ViL2 在医学图像分割领域展现出了显著的优势，但仍有提升空间。未来的研究将致力于增强模型的鲁棒性、扩展其应用范围以及优化其计算效率，以进一步推动医学图像分割技术的发展，为临床诊断和治疗提供更加精准和高效的工具。

参考文献

[1] Maximilian Beck, Korbinian Pöppel, Markus Spanring, Andreas Auer, Oleksandra Prudnikova, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. xlstm: Extended long short-term memory. In *Thirty-eighth Conference on Neural Information Processing Systems*, 2024.

[2] Matteo Cameli, Sergio Mondillo, Marco Solari, Francesca Maria Righini, Valentina Andrei, Carla Contaldi, Eugenia De Marco, Michele Di Mauro, Roberta Esposito, Sabina Gallina, et al. Echocardiographic assessment of left ventricular systolic function: from ejection fraction to torsion. *Heart Failure Reviews*, 21:77–94, 2016.

[3] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023.

[4] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.

[5] Jun Ma, Feifei Li, and Bo Wang. U-mamba: Enhancing long-range dependency for biomedical image segmentation, 2024.

[6] Keiron O'Shea and Ryan Nash. An introduction to convolutional neural networks, 2015.

[7] David Ouyang, Bryan He, Amirata Ghorbani, Matthew P. Lungren, Euan A. Ashley, David H. Liang, and James Y. Zou. Echonet-dynamic: a large new cardiac motion video data resource for medical machine learning. 2019.

[8] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[9] Alex Sherstinsky. Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *Physica D: Nonlinear Phenomena*, 404:132306, March 2020.

[10] Christian Bakke Vennerød, Adrian Kjærran, and Erling Stray Bugge. Long short-term memory rnn, 2021.