MemSAM-AMW：用于心脏超声视频分割的记忆分割模型

摘要

本文介绍了一种基于记忆机制的心脏超声视频分割方法（MemSAM），并在此基础上进一步引入了自适应记忆权重机制（Adaptive Memory Weighting, AMW），以提升模型在心脏超声视频分割任务中的性能。MemSAM 通过引入记忆模块来处理心脏超声视频中的时序信息，能够有效捕捉视频帧之间的长期依赖关系。然而，传统的记忆机制在更新过程中对所有帧的特征采用固定权重，忽略了不同帧的重要性差异。为了解决这一问题，本文提出了 AMW 机制，通过一个轻量级的注意力模块动态计算每一帧的重要性分数，并根据这些分数调整记忆更新时的权重。这种方法使模型能够更有效地利用信息丰富的帧（如心脏结构清晰的帧），同时减少对模糊或噪声帧的依赖。实验结果表明，MemSAM 结合 AMW 机制显著提高了模型的分割精度和鲁棒性。

关键词：心脏超声视频；有限标注；记忆读取；记忆更新；医学图像分割

1 引言

根据世界卫生组织（WHO）的统计数据，心血管疾病是全球死亡的首要原因 [7]。心脏超声图是评估心血管功能的一种重要而独特的工具。由于其便携性、低成本和实时性，在临床实践中的应用，心脏超声图通常被用作一线检查方法 [1]。然而，心脏超声图通常需要经验丰富的医生进行手动评估，评估的质量在很大程度上依赖于医生的专业知识 [18, 20]。最后，在人工评估中，观察者之间通常存在很大差异 [17]。此外，评估需要手动跟踪心室大小，这是费力、耗时和容易出错的。在这方面，在临床实践中高度需要自动化评估方法。

心脏超声图评估和诊断通常基于射血分数和心室容积的解释 [5]，这需要从心脏超声视频中准确分割关键结构，如左心室心内膜。然而，心脏超声图的自动分割一直以来都是一项具有挑战性的任务。首先，如图1(a, b) 所示，由于超声成像的局限性，存在很多影响心脏超声视频质量的不利因素，如低信噪比、斑点噪声、边缘丢失、致密肌肉和肋骨等结构造成的阴影，使得难以识别关键解剖结构的边界 [4, 30]。其次，视频内和视频间心脏结构的形状和比例变化较大（见图1(c, d)）。最后，心脏超声视频的注释是劳动密集型和耗时的，因此医师通常仅注释收缩末期舒张末期帧。所以我们必须分割具有有限和稀疏注释的心脏超声视频。

近年来，针对心脏超声视频分割提出了许多深度学习方法 [32–35]，但由于超声视频质量较低且标注有限，这些方法仍无法取得令人满意的结果。最近，已经提出了一种大的视觉模型，即分割任意对象模型（SAM）[16]，并且在许多自然图像分割任务中取得了显著的成功。

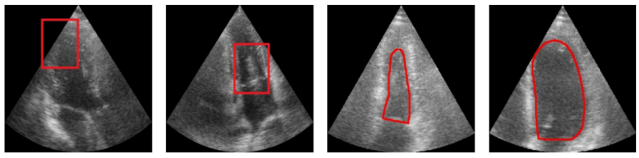

图 1. 心脏超声视频分割的挑战：(a) 轮廓模糊，(b) 斑点噪声，和 (c-d) 跨帧（同一视频的两帧）的比例变化。

一些研究者试图将其应用于医学图像分割任务，以利用 SAM 强大的表示能力来缓解训练样本不足的问题。但这些研究过于注重二维图像的分割，如何在医学视频分割中应用 SAM 仍然是一个尚未探索和具有挑战性的课题。将 SAM 直接应用于视频将忽略时间线索，可能导致时间上不一致的分割 [26,33]。例如，如图1所示，快速变化的心脏超声视频在目标对象的形状和尺度上具有明显的时间不连续性。此外，大量散斑噪声和伪影引起的模糊边界将极大地阻碍 SAM 释放其表示能力。本文以医学视频为研究对象设计了一种新的心脏超声视频分割模型，该模型具有医学视频与自然视频不同的特点。该模型的核心技术是一种具有时间感知能力和抗噪声能力的提示机制。具体地说，通过使用一个同时包含空间和时间信息的时空记忆来提示当前帧的分割，因此本文所提出的模型称为 MemSAM。在提示中，携带时间线索的记忆逐帧地顺序提示视频分割。同时，由于记忆提示传播了高层特征，避免了掩膜传播导致的误识别问题，提高了表示的一致性。为了解决散斑噪声的挑战，本文提出了一种记忆增强机制，该机制在存储之前利用预测掩码来提高记忆质量。在 SAMUS [19] 上构建了 MemSAM，SAMUS 是基于 SAM 的医学基础模型，这使得 MemSAM 以及 MemSAM-AMW 更适合于医学数据。最后，本文在两个公开数据集上进行了大量的实验。贡献可概括如下：

1) 介绍了一种新的基于 SAM 的心脏超声视频分割模型——MemSAM。模型的核心组件是一个新的提示方法，它能够提供空间和时间线索，以提高表示的一致性和分割精度。

2) 进一步提出了记忆增强模块，以在存储之前增强记忆，从而减轻记忆提示期间的斑点噪声和运动伪影的不利影响。

3) 引入了自适应记忆权重机制（AMW），通过轻量级的注意力模块动态计算每一帧的重要性分数，并根据这些分数调整记忆更新时的权重，使模型能够更有效地利用关键帧信息，减少对噪声帧的依赖。

4) 在两个公共数据集上广泛评估了 MemSAM-AMW，并与现有模型相比展示了最先进的性能。特别地，MemSAM-AMW 实现了与具有有限注释的完全监督方法相当的性能。

2 相关工作

2.1 医学图像分割中的 SAM

SAM 在应用于自然图像时表现出出色的零拍摄泛化能力 [12,23]。然而，由于医学图像中固有的复杂形状、模糊边界和显著尺度变化，它仍然不能直接应用于医学图像分割 [15]。一些一般性的工作试图将 SAM 从自然图像调整到医学图像 [41]，例如 MedSAM [22]，MSA [36]

和 SAMed [39]。MedSAM 并没有改变 SAM 网络结构，而是采用了更适合医学领域的边界框提示，并着重对掩码解码器进行了微调。MSA 和 SAMed 的目的是修改图像编码器以适应医学图像。MSA 通过向图像编码器添加适配器来实现这一点。SAMed 使用基于低秩（LoRA）策略的策略来微调图像编码器。在更专业的领域中，还提出了专注于超声图像的 SAMUS [19] 和 SonoSAM [28]。其中，SAMUS 通过添加适配器和额外的 CNN 分支更好地适应超声图像。SonoSAM 使用知识蒸馏从医学图像中提取特定知识。然而，这些方法仅限于图像分割，尚未扩展到视频数据，严重依赖于密集的注释和提示，以达到足够的性能。相比之下，这项工作的目的是研究利用视频中的时间线索，使模型训练只有稀疏的注释和最小的提示。

2.2 视频分割中的 SAM

虽然 SAM 的扩展到视频域仍然相对欠探索，一些初步的工作已经提出了解决自然的视频分割任务。一种常见的方法涉及将 SAM 与流行的视频分割架构集成，如 SAM-Track [10] 和 TAM [37] 所示。SAM-Track 使用 SAM 获取关键帧片段作为参考，然后利用 DeAOT [38] 在整个视频序列中传播参考帧。TAM 结合了 SAM 和 XMem [8]，首先生成带有 SAM 和弱提示的粗略掩码，然后使用 XMem 进行继续跟踪。当分割质量下降时，TAM 使用 XMem 的预测概率和亲和力作为提示来细化 SAM 输出。最近，SAM-PT [27] 引入了一种独特的点跟踪技术来生成掩码和跟踪对象。然而，这些方法仅适用于相对简单的自然图像场景，难以应用于医学图像分割。例如，对于复杂的、动态的和有噪声的超声图像，XMem 的中间特征将携带不正确地提示 SAM 的噪声。此外，当 XMem 将中间参数传递给 SAM 时，将它们转换为掩码提示符会丢失原始特性的高级语义。与 TAM 类似，我们的方法也基于 XMem。关键的是，我们的方法强调在特征转移期间保持语义一致性，并减轻医学成像数据中普遍存在的背景噪声，而不是简单的组合。

2.3 时空记忆方法

主流的视频时域建模方法包括多帧聚合和时空记忆网络。多帧聚合通过聚合相邻帧的语义信息来学习时间特征。相比之下，时空记忆通过沿着时间维度传播语义信息来对视频时间信息进行建模。虽然多帧聚合被广泛使用，但其 GPU 内存需求随着视频长度的增加而迅速增加，限制了其在长视频处理中的应用。相比之下，时空存储网络可以在确保时间建模的同时显著降低内存消耗，使其更适合扩展到医疗视频分析等领域。时空记忆网络（STM）首先由 Oh 等人 [24] 提出，用于视频对象分割任务。包括 STCN [9]、XMem [8] 和 XMem++ [3] 在内的后续方法已经证明了通用视频分割的巨大潜力。然而，这些方法需要一个带注释的参考关键帧来分割视频，这对于我们的任务来说是困难的。

3 本文方法

SAM 是一个强大的基于提示的分割框架，在学习了良好的表征之后，利用提示来跟踪分割的目标 [31,40]。如何恰当地进行提示是一个值得研究的问题. 现有的 SAM 及其变体在图像分割方面表现良好，包括自然图像和医学图像。然而，当直接迁移到视频分割时，它们无法利用视频中的时间线索，忽略了视频中的时空一致性。此外，将其直接应用于视频将需要提示每一帧，这对于视频分割来说是不优雅且冗余的。本文的目标是设计一种记忆提示方法来

扩展 SAM 框架，避免在视频中每一帧都提示。同时，对视频中的每一帧进行注释也是非常困难的，尤其是对于难以采集丰富的注释的心脏超声图。因此，需要一种能够完成半监督任务的方法。因此，本文提出 MemSAM 来解决心脏超声图中注释和提示较少的半监督问题。所提出的 MemSAM 框架以逐帧的顺序方式处理视频，如图2所示。长度为 T 帧的每个输入视频被一次一帧地馈送到 MemSAM 模型中。最初，提供前景中的随机采样点作为第一帧的提示，以引导模型。对于后续帧，MemSAM 仅依赖于记忆提示而不是外部提示。在 MemSAM 预测之后，受监督帧的预测将使用地面真实值计算损失。

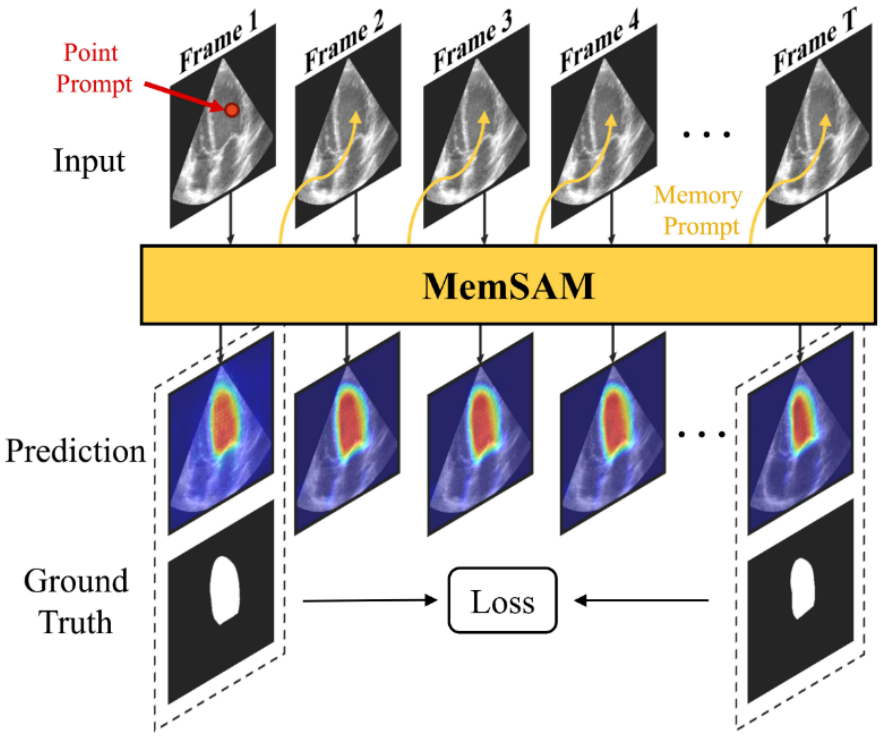

图 2. MemSAM 的工作流程，其中仅视频的第一帧使用最简单的正点提示（红色箭头），后续帧使用记忆提示（黄色箭头）。最后，计算监督帧的预测和真实值的损失。

3.1 本文方法概述

此部分对本文将要复现的工作进行概述，如图3所示：

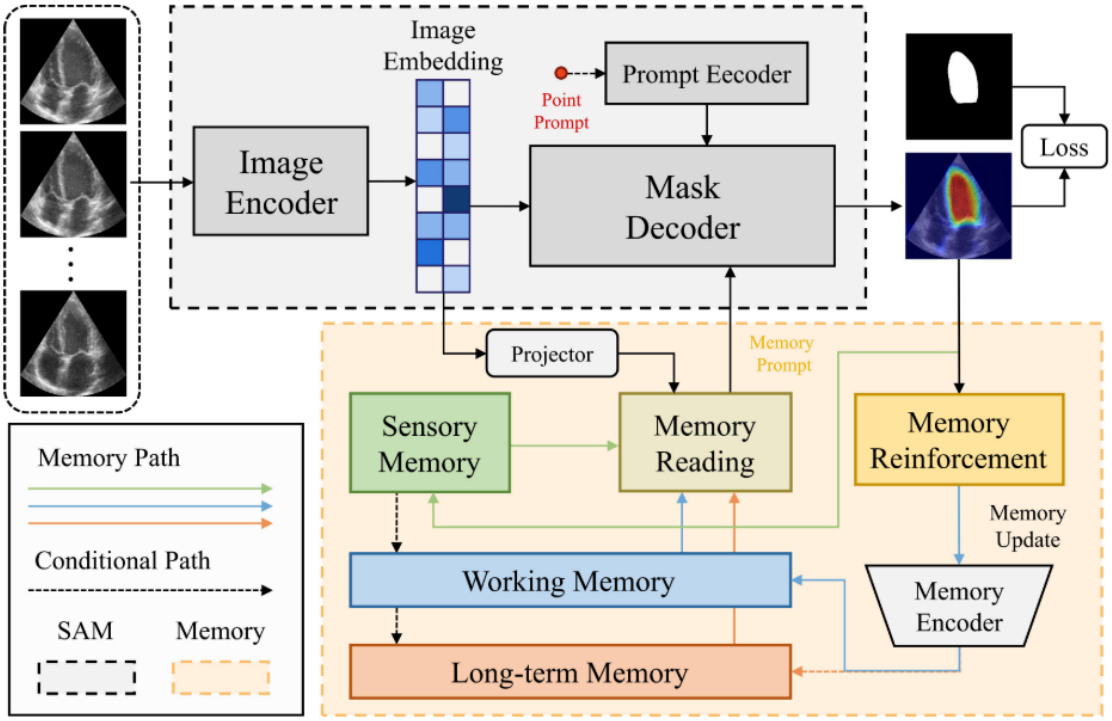

图 3. MemSAM 概述，由 SAM 和记忆组成。输入图像首先由 SAM 的图像编码器编码为图像向量。在获得点提示和记忆提示之后，从掩码解码器输出掩码。

图3显示了 MemSAM 框架内部的更多详细信息。MemSAM 主要由两个组件组成，SAM 组件和 Memory 组件。SAM 组件采用与原始 SAM 相同的架构，由图像编码器、提示编码器和掩码解码器组成。图像编码器采用 Vision Transformer（ViT）[2] 作为骨干，将输入图像编码为图像向量 E_i。提示编码器摄取外部提示，例如点提示，并将其编码为 c 维向量。随后，掩码解码器整合图像并提示向量以预测分割掩码。

其中，图像向量通过投影层映射到记忆特征空间，我们执行记忆阅读以从多特征记忆（感觉记忆、工作记忆和长期记忆）获得记忆提示，并将其提供给掩码解码器。最后，通过记忆加固和记忆编码器后，记忆将被更新。

3.2 记忆读取

图4中的记忆读取块示出了从图像向量 E_i 生成记忆向量 E_m 的过程，图像向量 E_i 作为记忆提示输入到掩码解码器。通过投影层投影帧 t 的图像向量 E_i^t 以生成查询 q^t。然后，该查询 q^t 用于针对记忆键和值执行亲和查询以获得读出特征 F^t。该过程可表述为:

$$F^t = v^{t-1} \cdot W(k^{t-1}, q^t) \tag{1}$$

其中，$k^{t-1} = k_w^{t-1} \oplus k_{lt}^{t-1}$，$v^{t-1} = v_w^{t-1} \oplus v_{lt}^{t-1}$，$\oplus$ 表示串联，下标 w 和 lt 分别表示工作记忆和长期记忆。$W(k^{t-1}, q^t)$ 表示查询 q^t 和记忆键 k^{t-1} 之间的关联矩阵，它捕获了 q^t 和 k^{t-1} 之间的相关性。它可以通过计算 q^t 和 k^{t-1} 之间的相似度，然后进行归一化来获得。具体计算过程可表述为:

$$W(k^{t-1}, q^t) = \text{softmax}(S(k^{t-1}, q^t)) \tag{2}$$

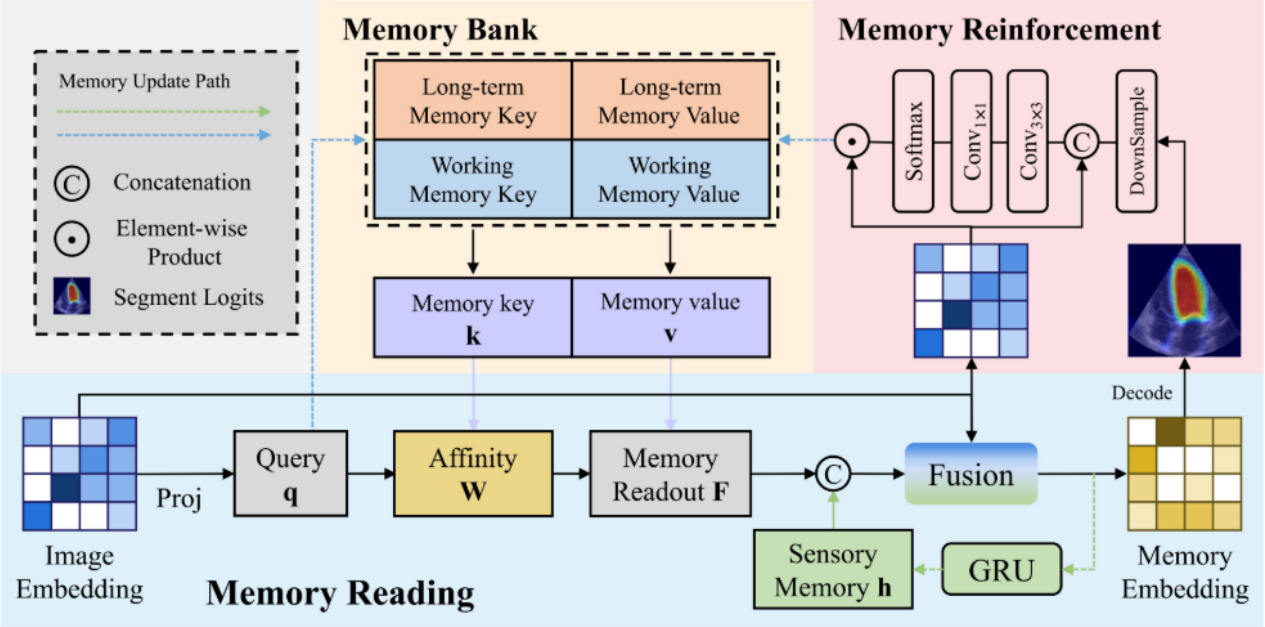

图 4. 更多关于记忆阅读和记忆强化的详细信息。

其中 S 是相似度计算。为了对记忆元素的置信水平进行编码并专注于更重要的通道，我们采用各向异性 L2 相似性 [8] 作为相似性函数。最后，读出特征 F^t 与感觉记忆 h^{t-1} 和 E_i^t 融合，以获得记忆向量 E_m^t，其公式为：

$$E_m^t = F\text{`aston}(E_i^t, F^t \oplus h^{t-1}) \tag{3}$$

3.3 记忆强化

与自然图像相比，超声图像中含有更为复杂的噪声，这意味着图像编码器产生的图像向量不可避免地会携带噪声。如果不经过任何处理就将噪声特征更新到记忆中，则可能导致错误的积累和传播。为了减轻噪声对记忆更新的影响，我们采用了记忆强化模块来增强记忆中特征表示的可辨别性。如图4中的记忆强化所示，我们在记忆更新之前用分割结果强化记忆，旨在强调前景特征并减少背景噪声的影响。具体来说，对于掩码解码器输出的概率图 $P^t \in \mathbb{R}^{B\times 1\times H\times W}$，我们首先将其下采样为与图像特征 $E_i^t \in \mathbb{R}^{B\times C\times h\times w}$ 相同大小的 $P_d^t \in \mathbb{R}^{B\times 1\times h\times w}$，然后将其与 E_i^t 沿通道维度拼接，得到 $F^t \in \mathbb{R}^{B\times (C+1)\times h\times w}$。我们使用一个卷积核大小为 3×3 的卷积层 $Conv_{3\times 3}$ 来处理 F，以限制每个像素的感受野。这一过程生成了局部注意力权重特征 $F_w^t \in \mathbb{R}^{B\times C_{mid}\times h\times w}$。接着，我们使用一个 $Conv_{1\times 1}$ 卷积层来改变 F_w^t 的通道数，得到 $F_w^t \in \mathbb{R}^{B\times C\times h\times w}$，最后计算输出特征：

$$F_o^t = P_d^t \odot \text{softmax}(F_w^t) \tag{4}$$

其中 $\odot$ 表示逐元素乘积。F_o 最终会被插入到工作记忆的值中。通过这种机制，我们利用分割结果来维护前景特征，减弱背景噪声对记忆更新的影响，并增强记忆中特征表达的可区分性。

3.4 记忆更新

需要更新的记忆包括记忆库和感官记忆。记忆库进一步由工作记忆和长期记忆组成，其中长期记忆仅用于长视频，此处省略。感官记忆 h^t 通过以下方式更新：

$$h^t = \mathrm{GRU}(h^{t-1} \oplus E_m^t) \tag{5}$$

其中 GRU 是门控循环单元 [11]。工作记忆的键 k_w^t 和值 v_w^t 更新如下：

$$k_w^t = q^t, \quad v_w^t = v_w^{t-1} \oplus F_\circ^t \tag{6}$$

4 复现细节

4.1 数据集介绍

在 CAMUS [17] 和 EchoNet-Dynamic [25] 这两个广泛使用的公开可用的心脏超声视频数据集上评估 MemSAM。

CAMUS 数据集包含 500 个病例，其中包括二维尖顶双腔和尖顶四腔视图视频。CAMUS 提供跨所有帧的注释。

EchoNet-Dynamic 数据集包含 10030 个二维尖顶双腔视图视频。每个视频都以积分的形式给出了左心室的面积。只标记收缩期和舒张末期。

为了全面评估 MemSAM 在半监督视频分割中的有效性，本文将 CAMUS 数据集改编为 CAMUS- full 和 CAMUS- semi 两种变体。CAMUS-Full 在训练期间对所有帧使用注释，而 CAMUS-Semi 仅对舒张末期 (ED) 和收缩末期 (ES) 帧使用注释。在测试期间，两个数据集都使用完整注释进行评估。从数据集中统一采样视频，将它们裁剪为 10 帧。裁剪确保 ED 帧是第一帧，ES 帧是最后一帧，分辨率调整为 256×256。对于 CAMUS 数据集，以 7:1:2 的比例将其分为训练集、验证集和测试集，而对于 EchoNet-Dynamic 数据集，使用了原始的分割方法。

4.2 与已有开源代码对比

在实现 MemSAM 模型的过程中，我主要参考了以下开源代码和研究成果进行实验：

1) 使用了 SAMUS [19] 作为 SAM 组件的实现基础。SAMUS 是针对超声图像优化的 SAM 模型，具有更友好的部署成本。本文继承了 SAMUS 的图像编码器结构，并对其各层进行了训练，其余部分则保留了原始 SAM 的参数并冻结。

2) 采用了 AdamW 优化器 [21] 进行模型训练，这是一种广泛使用的优化算法，能够有效处理权重衰减问题。

3) 使用了与 SAMUS 相同的损失函数，包括 Dice 损失 [13] 和二元交叉熵损失，这些损失函数在医学图像分割任务中表现良好。

4) 在训练阶段，应用了伽马增强、随机缩放、随机旋转和随机对比等数据增强技术，每种增强方法的概率为 0.5，以提高模型的泛化能力。

4.3 性能指标

本文采用了广泛使用的指标，如平均 Dice 系数（mDice）和平均交集（mIoU）进行分割评估，沿着 Hausdorff 距离-95%（HD 95）和平均对称表面距离（ASSD），报告了这些指标的标准差。

此外，本文还报告了左心室射血分数（LVEF）的三个统计指标。根据 CAMUS 数据集中提供的 Simpson 双平面圆盘法（SMOD）估计预测 LVEF。但需要注意的是，不同的实施方法将对最终 LVEF 结果产生显著影响。SMOD 根据心尖两腔和四腔视图的舒张末期和收缩末期时间实例估计 LVEF。与 Simpson 的单平面准则相比，SMOD 的估计解更精确、更可靠。对于预测和真实 LVEF，本文计算了 Pearson 相关系数（corr）、平均偏倚（bias）和标准误（std）。

4.4 创新点：基于自适应记忆权重的心脏超声视频分割

MemSAM 已经引入了记忆机制来处理心脏超声视频分割任务，但记忆更新过程中对所有特征的权重是固定的。然而，心脏超声视频中不同帧的重要性可能不同（例如，某些帧可能包含更清晰的心脏结构信息），因此固定的记忆权重可能无法充分利用关键帧的信息。

本文提出一种自适应记忆权重机制（Adaptive Memory Weighting, AMW），在记忆更新过程中动态调整不同帧特征的权重。具体来说：

1）通过一个轻量级的注意力模块，计算每一帧的重要性分数。

2）根据分数动态调整记忆更新时的权重，使模型更关注信息丰富的帧（如心脏结构清晰的帧），而减少对模糊或噪声帧的依赖。

在记忆更新模块中，增加一个轻量级的注意力层，计算每一帧的重要性分数 s^t：

$$s^t = \text{Sigmoid}(Conv_{1\times1}(E_i^t)) \tag{7}$$

使用分数 s^t 对记忆更新进行加权：

$$v_w^t = v_w^{t-1} \oplus (s^t \cdot F_o^t) \tag{8}$$

通过提出的 AMW 可以提高模型对关键帧的利用率，增强分割精度，并且减少噪声帧对记忆更新的干扰，提升模型的鲁棒性。

5 实验结果分析

5.1 与 SOTA 方法相比较

本文广泛地选择了不同类型的比较方法，包括传统的图像分割模型和医学基础模型。三种传统的图像分割模型分别是基于 CNN 的 UNet [29]，基于 Transformer 的 SwinUNet [6] 和 CNN-Transformer 混合 H2 Former [14]。医学适应 SAM 模型包括 MedSAM [22]、MSA [36]、SAMed [39]、SonoSAM [28] 和 SAMUS [19]。其中，SonoSAM 和 SAMUS 专注于超声图像。

定量比较结果如表 1 所示。在这些最先进的方法中，H2Former 和 SAMUS 在两个数据集上表现相对较好，分别受益于 CNN-Transformer 架构和超声图像优化。然而，如果没有利用稀

缺注释下视频的时间属性，这些模型仍然滞后于 MemSAM 以及本文的创新模型 MemSAM-AMW。实验验证了我们的方法在给定有限注释的情况下达到了最先进的性能。

表 1. 在 CAMUS-Semi 和 EchoNet-Dynamic 数据集上，采用最先进的方法对所提出的方法进行分割的性能。HD95 和 ASSD 在 CAMUS-Semi 中以毫米 (mm) 为单位测量，而在 EchoNet-Dynamic 中以像素为单位测量。我们的结果表示为平均值 ± 标准差。

Method	CAMUS - Semi				EchoNet - Dynamic			
	mDice ↑	mIoU ↑	HD95 ↓	ASSD ↓	mDice ↑	mIoU ↑	HD95 ↓	ASSD ↓
UNet [29]	90.13	82.36	5.77	2.35	91.36	83.27	4.98	3.01
SwimUNet [6]	88.84	80.33	6.10	2.60	87.79	80.14	6.61	5.71
H2Former [14]	91.31	84.30	5.27	2.05	90.21	82.46	5.12	3.78
MedSAM [22]	85.42	75.14	8.42	3.34	86.47	79.19	7.97	4.88
MSA [36]	88.03	78.98	7.53	2.85	87.91	78.34	6.67	4.34
SAMed [39]	87.45	78.14	9.17	3.10	86.35	78.96	7.12	4.59
SonoSAM [28]	89.80	81.79	6.60	2.45	89.61	82.33	6.58	3.80
SAMUS [19]	91.11	83.94	5.08	2.07	91.79	84.32	5.35	3.22
MemSAM	93.31	87.61	3.82	1.57	92.78	85.89	4.57	2.71
MemSAM-AMW	**93.51** ± 1.04	**87.91** ± 4.12	**3.92** ± 1.90	**1.87** ± 0.22	**92.88** ± 3.48	**85.99** ± 5.22	**4.77** ± 2.44	**2.81** ± 0.98

我们的方法与最先进的 LVEF 估计方法的比较如表 2 所示。在有限的标注条件下，现有的最先进的方法在准确率上并不令人满意，这主要归因于两个因素。首先，最先进的方法本身的分割精度仍然不足。其次，SMOD 估计方案要求高分割质量，需要两室和四室视图来产生准确的量化，以进行鲁棒的 LVEF 评估。

表 2. CAMUS-Semi 数据集上不同临床指标的比较

Method	CAMUS-Semi		
	corr (%) ↑	bias ↓	std ↓
UNet [29]	67.15	11.65	9.39
SwimUNet [6]	59.41	6.90	9.06
H2Former [14]	58.61	**0.69**	**7.49**
MedSAM [22]	41.63	11.22	11.19
MSA [36]	31.00	13.25	14.96
SAMed [39]	28.22	13.34	12.24
SonoSAM [28]	56.18	11.83	9.12
SAMUS [19]	67.55	7.02	9.16
MemSAM	78.92	4.86	11.10
MemSAM-AMW	**78.99**	4.57	10.90

6 总结与展望

本文的主要介绍了 MemSAM 框架，并通过引入自适应记忆权重机制（AMW）进一步优化了其性能。MemSAM 通过记忆模块有效捕捉了心脏超声视频中的时序信息，解决了传统方法在处理长时序数据时的局限性。然而，MemSAM 在记忆更新过程中对所有帧的特征采用固

定权重，未能充分利用关键帧的信息。为此，本文提出了 AMW 机制，通过轻量级的注意力模块动态计算每一帧的重要性分数，并根据这些分数调整记忆更新时的权重。AMW 机制使模型能够更关注信息丰富的帧，减少对噪声帧的依赖，从而显著提升了分割精度和鲁棒性。实验结果表明，MemSAM 结合 AMW 机制在心脏超声视频分割任务中表现出色，为未来的医学图像分割研究提供了新的思路。未来的工作可以进一步探索 MemSAM 和 AMW 机制在其他医学图像分割任务中的应用，以验证其通用性和扩展性。

参考文献

[1] Zeynettin Akkus, Yousof H Aly, Itzhak Z Attia, Francisco Lopez-Jimenez, Adelaide M Arruda-Olson, Patricia A Pellikka, Sorin V Pislaru, Garvan C Kane, Paul A Friedman, and Jae K Oh. Artificial intelligence (ai)-empowered echocardiography interpretation: a state-of-the-art review. *Journal of clinical medicine*, 10(7):1391, 2021.

[2] Dosovitskiy Alexey. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv: 2010.11929*, 2020.

[3] Maksym Bekuzarov, Ariana Bermudez, Joon-Young Lee, and Hao Li. Xmem++: Production-level video segmentation from few annotated frames. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 635–644, 2023.

[4] Nagashettappa Biradar, Mohan Lal Dewal, and Manoj Kumar Rohit. Speckle noise reduction in b-mode echocardiographic images: A comparison. *IETE Technical Review*, 32(6):435–453, 2015.

[5] 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.

[6] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pages 205–218. Springer, 2022.

[7] Chen Chen, Chen Qin, Huaqi Qiu, Giacomo Tarroni, Jinming Duan, Wenjia Bai, and Daniel Rueckert. Deep learning for cardiac image segmentation: a review. *Frontiers in cardiovascular medicine*, 7:25, 2020.

[8] Ho Kei Cheng and Alexander G Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *European Conference on Computer Vision*, pages 640–658. Springer, 2022.

[9] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. *Advances in Neural Information Processing Systems*, 34:11781–11794, 2021.

[10] Yangming Cheng, Liulei Li, Yuanyou Xu, Xiaodi Li, Zongxin Yang, Wenguan Wang, and Yi Yang. Segment and track anything. *arXiv preprint arXiv:2305.06558*, 2023.

[11] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[12] C Cui, R Deng, Q Liu, T Yao, S Bao, LW Remedios, Y Tang, and Y Huo. All-in-sam: From weak annotation to pixel-wise nuclei segmentation with prompt-based finetuning. arxiv. *arXiv preprint arXiv:2307.00290*, 2023.

[13] N Fausto Milletari and Ahmadi Seyed-Ahmad V-Net. Fully convolutional neural networks for volumetric medical image segmentation.

[14] Along He, Kai Wang, Tao Li, Chengkun Du, Shuang Xia, and Huazhu Fu. H2former: An efficient hierarchical hybrid transformer for medical image segmentation. *IEEE Transactions on Medical Imaging*, 42(9):2763–2775, 2023.

[15] Yuhao Huang, Xin Yang, Lian Liu, Han Zhou, Ao Chang, Xinrui Zhou, Rusi Chen, Junxuan Yu, Jiongquan Chen, Chaoyu Chen, et al. Segment anything model for medical images? *Medical Image Analysis*, 92:103061, 2024.

[16] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[17] Sarah Leclerc, Erik Smistad, Joao Pedrosa, Andreas Østvik, Frederic Cervenansky, Florian Espinosa, Torvald Espeland, Erik Andreas Rye Berg, Pierre-Marc Jodoin, Thomas Grenier, et al. Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. *IEEE transactions on medical imaging*, 38(9):2198–2210, 2019.

[18] Honghe Li, Yonghuai Wang, Mingjun Qu, Peng Cao, Chaolu Feng, and Jinzhu Yang. Echoefnet: multi-task deep learning network for automatic calculation of left ventricular ejection fraction in 2d echocardiography. *Computers in Biology and Medicine*, 156:106705, 2023.

[19] Xian Lin, Yangyang Xiang, L Zhang, X Yang, Z Yan, and L Yu. Samus: adapting segment anything model for clinically-friendly and generalizable ultrasound image segmentation (2023). *arXiv preprint arXiv:2309.06824*.

[20] Fei Liu, Kun Wang, Dan Liu, Xin Yang, and Jie Tian. Deep pyramid local attention neural network for cardiac structure segmentation in two-dimensional echocardiography. *Medical image analysis*, 67:101873, 2021.

[21] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[22] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024.

[23] S Mo and Y Tian. Av-sam: Segment anything model meets audio-visual localization and segmentation. arxiv 2023. *arXiv preprint arXiv:2305.0183*.

[24] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9226–9235, 2019.

[25] David Ouyang, Bryan He, Amirata Ghorbani, Matt 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. In *NeurIPS ML4H Workshop*, pages 1–11, 2019.

[26] Nathan Painchaud, Nicolas Duchateau, Olivier Bernard, and Pierre-Marc Jodoin. Echocardiography segmentation with enforced temporal consistency. *IEEE Transactions on Medical Imaging*, 41(10):2867–2878, 2022.

[27] Frano Rajič, Lei Ke, Yu-Wing Tai, Chi-Keung Tang, Martin Danelljan, and Fisher Yu. Segment anything meets point tracking. *arXiv preprint arXiv:2307.01197*, 2023.

[28] Hariharan Ravishankar, Rohan Patil, Vikram Melapudi, and Pavan Annangi. Sonosamsegment anything on ultrasound images. In *International Workshop on Advances in Simplifying Medical Ultrasound*, pages 23–33. Springer, 2023.

[29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[30] Ayesha Saadia and Adnan Rashdi. A speckle noise removal method. *Circuits, Systems, and Signal Processing*, 37:2639–2650, 2018.

[31] Xuemeng Song, Liqiang Jing, Dengtian Lin, Zhongzhou Zhao, Haiqing Chen, and Liqiang Nie. V2p: Vision-to-prompt based multi-modal product summary generation. In *Proceedings of the 45th International ACM SIGIR conference on research and development in information retrieval*, pages 992–1001, 2022.

[32] S Thomas, A Gilbert, and G Ben-Yosef. Light-weight spatio-temporal graphs for segmentation and ejection fraction prediction in cardiac ultrasound, 380–390 (2022). *DOI: https://doi. org/10.1007/9783031164408_37*.

[33] Hongrong Wei, Heng Cao, Yiqin Cao, Yongjin Zhou, Wufeng Xue, Dong Ni, and Shuo Li. Temporal-consistent segmentation of echocardiography with co-learning from appearance and shape. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part II 23*, pages 623–632. Springer, 2020.

[34] Huisi Wu, Jingyin Lin, Wende Xie, and Jing Qin. Super-efficient echocardiography video segmentation via proxy-and kernel-based semi-supervised learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2803–2811, 2023.

[35] Huisi Wu, Jiasheng Liu, Fangyan Xiao, Zhenkun Wen, Lan Cheng, and Jing Qin. Semi-supervised segmentation of echocardiography videos via noise-resilient spatiotemporal semantic calibration and fusion. *Medical Image Analysis*, 78:102397, 2022.

[36] Junde Wu, Wei Ji, Yuanpei Liu, Huazhu Fu, Min Xu, Yanwu Xu, and Yueming Jin. Medical sam adapter: Adapting segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.12620*, 2023.

[37] Jinyu Yang, Mingqi Gao, Zhe Li, Shang Gao, Fangjing Wang, and Feng Zheng. Track anything: Segment anything meets videos. arxiv 2023. *arXiv preprint arXiv:2304.11968*, 2023.

[38] Zongxin Yang and Yi Yang. Decoupling features in hierarchical propagation for video object segmentation. *Advances in Neural Information Processing Systems*, 35:36324–36336, 2022.

[39] Kaidong Zhang and Dong Liu. Customized segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.13785*, 2023.

[40] Rongzhi Zhang, Yue Yu, Pranav Shetty, Le Song, and Chao Zhang. Prboost: Prompt-based rule discovery and boosting for interactive weakly-supervised learning. *arXiv preprint arXiv:2203.09735*, 2022.

[41] Yichi Zhang and Rushi Jiao. How segment anything model (sam) boost medical image segmentation: A survey. *Available at SSRN 4495221*, 2023.