arXiv:2305.01836v1 [cs.CV] 3 May 2023

AV-SAM：Segment Anything Model 结合音频-视觉定位与分割

Shentong Mo 卡内基梅隆大学

Yapeng Tian*德克萨斯大学达拉斯分校

摘要

Segment Anything Model（*SAM*）近期在视觉分割任务中展现出强大的效果。然而，对于 *SAM* 如何应用于音频-视觉任务（例如视觉声音定位与分割），目前的探索还不多。在这项工作中，我们提出了一种基于 *Segment Anything Model* 的简洁而有效的音频-视觉定位与分割框架，即 *AV-SAM*，它能够生成与音频对应的发声物体掩码。具体而言，*AV-SAM* 仅利用 *SAM* 预训练图像编码器输出的音频特征和视觉特征之间的逐像素音视融合来聚合跨模态表征。随后，将聚合后的跨模态特征输入提示编码器和掩码解码器，生成最终的音频-视觉分割掩码。我们在 *Flickr-SoundNet* 和 *AVSBench* 数据集上进行了大量实验。结果表明，所提出的 *AV-SAM* 在发声物体定位与分割方面取得了有竞争力的性能。

1. 引言

当我们在动物园听到狮子咆哮时，由于现实世界中音频信号与发声物体之间存在很强的相关性，我们自然能够意识到狮子所在的位置。这种源于人类的智能感知能力吸引了许多研究人员，促使他们探索联合音频-视觉学习，以实现视觉声音定位与分割。

音频-视觉定位与分割是一项颇具挑战的任务，旨在预测视频中发声物体的逐像素掩码。为了解决这一任务，近期的方法 [5, 6] 手动设计了多种对比学习流程，并以音频-视觉对齐为线索来定位声源。最近，AVGN [8]利用结合类别级语义的跨模态分组，从音频混合信号中预测各个声源。在这项工作中，我们将基于 Segment Anything Model，通过逐像素音视融合提取紧凑特征来解决这一问题。

*通讯作者。

Segment Anything Model（SAM） [4] 近期在图像分割任务中展现出卓越的效果。这个可通过提示进行操作的基础模型使用来自1100万张图像的10亿个掩码进行训练，由三个主要组件构成：图像编码器、提示编码器和掩码解码器。然而，目前尚无研究探索 SAM 如何应用于具有挑战性的音频-视觉任务，包括视觉发声定位与分割。

主要挑战在于，音频并不会与视频中存在的所有物体自然对齐。这促使我们从视频中为每个掩码提示学习与音频对齐的视觉特征，以指导 SAM 中的掩码生成。为了解决这一问题，我们的核心思路是利用音频特征与视觉特征之间的逐像素音视融合，聚合跨模态表征，并将其作为音视掩码解码器的输入；与以往的音频-视觉定位和分割方法相比，这种方法更加简洁。

为此，我们提出了一种新颖的、采用逐像素音视融合的 Segment Anything Model，即 AV-SAM，能够预测与音频信号相对应的发声物体掩码。具体而言，我们的 AV-SAM 通过显式的逐像素音视融合，聚合音频表征与 SAM 预训练图像编码器输出的视觉特征之间的跨模态表征。随后，音视掩码解码器将聚合后的跨模态特征和提示嵌入作为输入，生成最终的分割掩码。

我们在 Flickr-SoundNet 和 AVSBench 数据集上进行了大量实验。实验结果表明，所提出的 AV-SAM 在发声物体定位与分割方面能够取得与 SAM 相当的性能。此外，定性掩码可视化结果生动地展示了我们的 AV-SAM 在音频-视觉定位与分割方面的优势。

2. 相关工作

音频-视觉联合学习。 许多先前的研究都探讨了音频-视觉联合学习，旨在从视频中学习两种不同模态之间的音频-视觉对齐，以应用于多种音频-视觉任务，例如音频-视觉空间化 [9, 10]，音频事件定位 [13]

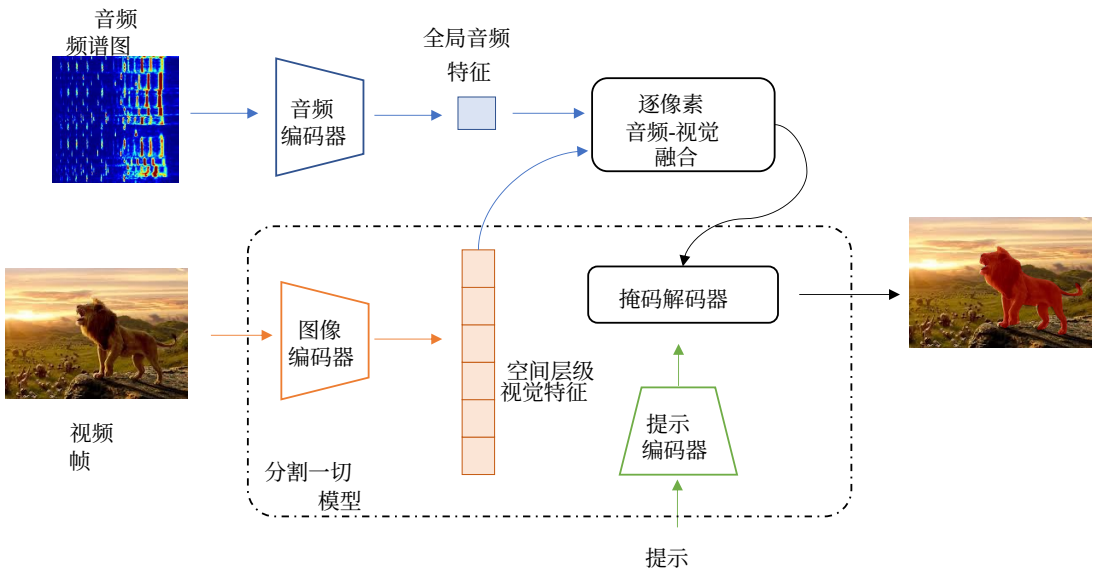

图1. 所提出的用于音频-视觉定位与分割的 Segment Anything Model（AV-SAM）示意图。逐像素音视融合模块聚合来自预训练音频和图像编码器的音频特征与视觉特征之间的跨模态表征。随后，聚合后的跨模态特征与提示编码器生成的提示嵌入被输入掩码解码器，以生成最终的分割掩码。

和音频-视觉解析 [7, 12]。本文主要关注学习用于音频-视觉定位与分割的判别性跨模态表征，这比上述任务更具挑战性。

音频-视觉定位与分割。 音频-视觉定位与分割是一项典型且具有挑战性的问题，旨在以热图或掩码的形式预测视频中各个声源的位置。近年来，研究人员 [5, 6, 8, 14] 探索了多种人工设计的流程，以学习音频与视觉之间的对应关系，从而定位单个或多个声源。与这些方法不同，我们提出了一种基于最新视觉基础模型 Segment Anything Model 的简单框架，同时实现音频-视觉定位与分割。

Segment Anything Model。 Segment Anything Model（SAM） [4] 是一种可提示模型，使用来自1100万张图像的 10 亿个掩码进行训练，用于图像分割。这个视觉基础模型由三个主要组件构成：图像编码器、提示编码器和掩码解码器。尽管该模型展示了出色的视觉分割效果，但 SAM 在音频-视觉定位与分割上的表现尚不明确。相比之下，我们设计了一种基于 SAM 的新颖框架，通过像素级音频-视觉融合来聚合音频和视觉模态之间的跨模态表征。

3. 方法

给定一幅图像和音频，我们的目标是预测图像中发声物体的逐像素掩码。我们提出一种名为 AV-SAM 的新型 Segment Anything Model，采用音视融合技术，用于视觉声音定位与分割，主要由两个模块组成：逐像素音视融合3.2 和第3.3节中的音视掩码解码器。

3.1. 预备知识

本节首先介绍问题设定和符号，然后回顾 SAM [4] 用于图像分割的掩码生成过程。

问题设定与符号。给定音频频谱图和一幅图像，我们的目标是预测图像中各个声源的空间位置。设$\mathcal{D} = (a_i, v_i) : i = 1, ..., N$是由音频 $a_i \in \mathrm{R}^{T\times F}$与视觉 $v_i \in \mathrm{R}^{3\times H\times W}$ 配对构成的数据集。注意， T, F 分别表示音频频谱图的时间维度和频率维度。遵循先前工作 [6, 14], , 我们首先使用双流编码器和投影头对音频与视觉输入进行编码，分别记为音频和图像对应的 $f_a(\cdot)$ $g_a(\cdot)$以及 $f_v(\cdot)$ $g_v(\cdot)$。音频编码器计算全局音频特征 $\mathrm{a}_i = g_a(f_a(a_i))$，其维度为 $\mathrm{a}_i \in \mathrm{R}^{1\times D}$；视觉编码器则为每个第$s$阶段生成多尺度空间层级特征 $\{\mathrm{v}_i^s\}_{s=1}^S = g_v(f_v(v_i))$，其维度为 $\mathrm{v}_i^s \in \mathrm{R}^{D\times H^s\times W^s}$。

重新审视 SAM。 为解决视觉分割问题，SAM [4] 引入了提示编码器 $f_p(\cdot)$和轻量级掩码解码器 $f_m(\cdot)$来生成分割掩码。具体而言，提示编码器以位置嵌入、稀疏提示类型（如点、框和文本）以及基于卷积的稠密掩码提示嵌入为输入

随后，掩码解码器应用了一个包含自注意力/交叉注意力的 Transformer 解码器块，并通过动态掩码预测头预测最终掩码。

3.2. 逐像素音视融合

为解决视听分割问题，我们受到 AVS 的启发 [14]，并引入逐像素音视融合模块，对多尺度空间级视觉特征和全局音频表征进行编码，从而更新输入，以供后续掩码解码器使用。经过跨模态融合后，第$\mathbf{z}_i^s$阶段的音频-视觉特征 s更新如下：

$$\mathbf{z}_i^s = \mathbf{v}_i^s + \mu\left(\frac{\theta(\mathbf{v}_i^s)\phi(\hat{\mathbf{a}}_i)^\top}{H^s W^s}\omega(\mathbf{v}_i^s)\right) \tag{1}$$

其中，$\hat{\mathbf{a}}_i \in \mathrm{R}^{D\times H^s\times W^s}$ 表示全局音频表征$\mathbf{a}_i$ 的复制版本，该版本在第 s阶段重复 $H^s W^s$ 次。这里， $\mu(\cdot)$ $\theta(\cdot)$ $\phi(\cdot)$ $\omega(\cdot)$表示 $1\times1\times1$ 卷积算子。通过这种显式的音视融合，我们将促使学得的视觉词元嵌入与全局音频特征实现有区分性的对齐。

3.3. 音视掩码解码器

得益于逐像素音视融合，我们将 SAM [4] 中预训练图像编码器输出的原始视觉特征替换为最后阶段的多尺度特征图$\mathbf{z}_i^s$。随后，将这些更新后的多阶段特征图输入 SAM [4] 中的掩码解码器和提示编码器，以生成最终输出掩码$\mathbf{M} \in \mathrm{R}^{H\times W}$。以像素级标注$\mathbf{Y}$ 作为监督，我们只需对预测结果和标签应用二元交叉熵（BCE）损失，如下所示：

$$\mathcal{L} = \mathrm{BCE}(\mathbf{M}, \mathbf{Y}) \tag{2}$$

在推理过程中，我们遵循先前的工作 [4]，并预测最终的音频-视觉分割掩码以进行评估。

4. 实验

4.1. 实验设置

数据集。 我们使用 VGG-Sound [1]中的 14.4 万对数据子集进行训练，并在 Flickr-SoundNet 测试集 [11] 上测试模型，该测试集包含 250 对发声物体的音频-视觉数据。需要注意的是，我们使用 ImageNet [2] 预训练的 ResNet50 [3]，通过对特征图进行双线性插值来生成伪掩码。AVSBench [14] 包含来自 23 个类别的 4,932 个视频（总计 10,852 帧），类别包括乐器、人类、动物等。遵循先前的工作 [14], ，我们使用相同的数据划分：3,452/740/740 个视频分别用于训练、验证和测试。

评估指标。 对于视觉声音定位，我们遵循先前的方法 [6, 8]，并使用像素级平均精度（AP）、交并比（IoU）和曲线下面积（AUC）进行评估。

方法	Flickr-SoundNet			AVS-Bench	
	AP(%)	IoU(%)	AUC(%)	mIoU(%)	F1(%)
SAM [4]	20.02	41.60	20.46	29.67	42.06
AV-SAM	**28.93**	**50.80**	**29.76**	**40.47**	**56.57**

表 1. Flickr-SoundNet 和 AVS-Bench 数据集上单源定位与分割的定量结果。

Mask 解码器	提示编码器	图像编码器	mIoU(%)	F1(%)
✗	✗	✗	32.58	45.36
✓	✗	✗	34.11	49.67
✓	✓	✗	36.63	51.58
✓	✓	✓	**40.47**	**56.57**

表 2. 针对 SAM 各模块在音频-视觉定位与分割中的冻结（）和微调（✓）进行的探索性研究。

对于音频-视觉分割，我们遵循先前的工作 [14]，采用 mIoU 和 F1 分数作为评估指标。

实现细节。 对于输入视觉帧，分辨率调整为 1024×1024。对于输入音频，我们使用长度为 $3s$ 、采样率为 22050Hz 的对数频谱图。遵循先前工作 [6], , 我们应用 STFT，通过使用 50ms 窗口和 25ms 步长，生成形状为 257×300 的输入张量（即，覆盖 300 个时间步的 257 个频率带）。我们使用轻量级 ResNet 18 [3] 作为音频编码器，并使用 SAM [4] 发布的权重初始化视觉模型。模型使用批量大小为 128 的 Adam 优化器训练 100 个周期，学习率为 $1e-4$。

4.2. 与先前工作的比较

在这项工作中，我们提出了一种基于 SAM [4] 的新颖且有效的框架，用于音频-视觉定位与分割。为验证所提出的 AV-SAM 的有效性，我们将其与视觉基础模型基线进行了全面比较。

对于 Flickr-SoundNet 上的视觉声音定位，我们在表 1 中报告了定量比较结果。可以看出，与 SAM [4], 这一有竞争力的视觉基础模型基线相比，我们在两个基准的所有指标上均取得了最佳结果。具体而言，所提出的 AV-SAM 在 Flickr-SoundNet 数据集上以 8.91 AP、9.20 IoU 和 9.30 AUC 的优势超越了 SAM [4]。此外，我们的 mIoU 和 F1 分别提升了 10.80 和 14.51，这表明逐像素音视融合对于聚合跨模态输入、为掩码解码器提供指导以生成具有区分性的分割掩码至关重要。这些显著提升证明了我们的方法在音频-视觉定位与分割方面优于 SAM [4]。

为了定性评估分割掩码，我们将所提出的 AV-SAM 与 SAM 进行比较 [4]，如图 2 所示。请注意，我们仅可视化一个



图 2. 与 SAM 在视听分割任务上的定性比较。我们提出的 AV-SAM 能为发声源生成更准确、质量更高的分割掩码。

由 SAM 合并得到的掩码 [4]，其置信度分数较高。通过比较可以观察到，在没有显式音频-视觉融合的情况下，SAM [4], 这一强大的视觉分割基线在发声物体上的表现较差。例如，给定一张女孩拉小提琴的图像，该基线模型倾向于预测同时覆盖女孩和小提琴的掩码。相比之下，我们的方法生成的分割掩码质量明显更高，也更接近真实标注。这些可视化结果进一步展示了我们简单的 AV-SAM 在基于 SAM 并采用显式逐像素音视融合进行视觉声音分割方面的优越性 [4]。

4.3. 实验分析

在本节中，我们进行了消融研究，以展示冻结和微调来自 SAM 的预训练权重所产生的影响 [4]。为了更全面地探究这一影响，我们在表中分别消融了对各模块（掩码解码器、提示编码器、图像编码器）参数进行微调的必要性2。我们可以观察到，在基础基线模型上微调掩码解码器可提升视听分割结果（mIoU 提高 1.53，F1 提高 4.31），这表明视听掩码解码器能够从聚合的跨模态特征中生成准确掩码。同时，微调提示编码器也提升了视觉声源分割在所有指标上的表现。更重要的是，在基线模型中同时微调全部三个模块，可使性能大幅提升，mIoU 提高 7.89，F1 提高 11.21。这些改进结果验证了在视听分割任务中，对 SAM 的预训练权重进行微调对于从音频和图像中提取跨模态语义的重要性。

5. 结论

在这项工作中，我们提出了 AV-SAM，这是一种基于 SAM 的简单而有效的音频-视觉框架，能够生成图像中的发声物体掩码。我们引入逐像素音视融合，以聚合音频与视觉特征之间的跨模态表征。随后，我们利用聚合后的跨模态特征，通过提示编码器和掩码解码器生成最终的音频-视觉分割掩码。在 Flickr-SoundNet 和 AVSBench 数据集上的实证实验表明，所提出的 AV-SAM 在视觉声音定位和分割方面表现出色。

参考文献

[1] Honglie Chen、Weidi Xie、Andrea Vedaldi 和 Andrew Zisserman。VggSound：一个大规模音频-视觉数据集。载于 ICASSP，2020。3[2] Jia Deng、Wei Dong、Richard Socher、Li-Jia Li、Kai Li 和 Li Fei-Fei。ImageNet：一个大规模分层图像数据库。载于 CVPR，2009。3[3] Kaiming He、Xiangyu Zhang、Shaoqing Ren 和 Jian Sun。用于图像识别的深度残差学习。载于 CVPR，第 770–778 页。3[4] Alexander Kirillov、Eric Mintun、Nikhila Ravi、Hanzi Mao、Chloe Rolland、Laura Gustafson、Tete Xiao、Spencer Whitehead、Alexander C. Berg、Wan-Yen Lo、Piotr Doll´ar 和 Ross Girshick。分割一切。arXiv:2304.02643，2023。1、2、3、4[5] Shentong Mo 和 Pedro Morgado。深入探究弱监督音频-视觉声源定位。载于 NeurIPS，2022。1、2[6] Shentong Mo 和 Pedro Morgado。轻松定位视觉声音。载于 ECCV，第 218–234 页，2022。1、2、3[7] Shentong Mo 和 Yapeng Tian。用于弱监督音频-视觉视频解析的多模态分组网络。载于 NeurIPS，2022。2[8] Shentong Mo 和 Yapeng Tian。用于混合音源定位的音频-视觉分组网络。arXiv 预印本 arXiv:2303.17056，2023。1、2、3[9] Pedro Morgado、Yi Li 和 Nuno Nvasconcelos。从音频-视觉空间对齐中学习表征。载于 NeurIPS，第 4733–4744 页，2020。1[10] Pedro Morgado、Nuno Nvasconcelos、Timothy Langlois 和 Oliver Wang。为 360° 视频自监督生成空间音频。载于 NeurIPS，2018。1[11] Arda Senocak、Tae-Hyun Oh、Junsik Kim、Ming-Hsuan Yang 和 In So Kweon。学习在视觉场景中定位声源。载于 CVPR，2018。3[12] Yapeng Tian、Dingzeyu Li 和 Chenliang Xu。统一的多感官感知：弱监督音频-视觉视频解析。载于 ECCV，第 436–454 页，2020。2[13] Yapeng Tian、Jing Shi、Bochen Li、Zhiyao Duan 和 Chenliang Xu。无约束视频中的音频-视觉事件定位。载于 ECCV，2018。1[14] Jinxing Zhou、Jianyuan Wang、Jiayi Zhang、Weixuan Sun、Jing Zhang、Stan Birchfield、Dan Guo、Lingpeng Kong、Meng Wang 和 Yiran Zhong。视听分割。载于 ECCV，2022。2、3