arXiv:2603.08967v1 [cs.CV] 9 Mar 2026

你能持续听辨、定位和分割吗？音视频分割无示例持续学习基准

Siddeshwar Raghavan[1], Gautham Vinod[1], Bruce Coburn[1], 和 Fengqing Zhu[1]

美国印第安纳州西拉斐特市普渡大学，邮编 47906
{raghav12, gvinod, coburn6, zhu0}@purdue.edu

摘要。 音视频分割（AVS）旨在通过联合学习音频和视觉信号，为视频中的发声物体生成像素级掩码。然而，现实环境本质上是动态的，音频和视觉分布会随时间不断演变，这给假设训练设置静态不变的现有音视分割系统带来了挑战。为弥补这一不足，我们提出首个音视频分割无示例持续学习基准，涵盖单源和多源音视分割数据集上的四种学习协议。此外，我们提出强大的基线方法 ATLAS，它通过投影后的音频上下文，在跨模态注意力之前采用音频引导的预融合条件化来调制视觉特征通道。最后，我们引入低秩锚定（LRA），根据损失敏感度稳定适配后的权重，从而缓解灾难性遗忘。大量实验表明，该方法在多种持续学习场景中表现出竞争力，为终身音视感知奠定了基础。代码见[*1] - https://gitlab.com/viper-purdue/atlas

关键词： 持续学习 · 音视频分割 · 多模态学习

1 引言

人类能够凭借日常经验，自然而然地结合视觉线索与物体发出的声音来识别并定位物体 [25, 31, 34]。这种能力是我们感知、理解复杂环境并与之互动的基础。音视频分割（AVS） [50, 51] 旨在应对这一挑战，为视频帧中发出声音的物体生成像素级映射。然而，人类会不断适应变化的环境，在获取新的视觉和听觉模式的同时保留先前学到的知识。比如，第一次听到一种新的乐器，并

[1] 论文审稿中

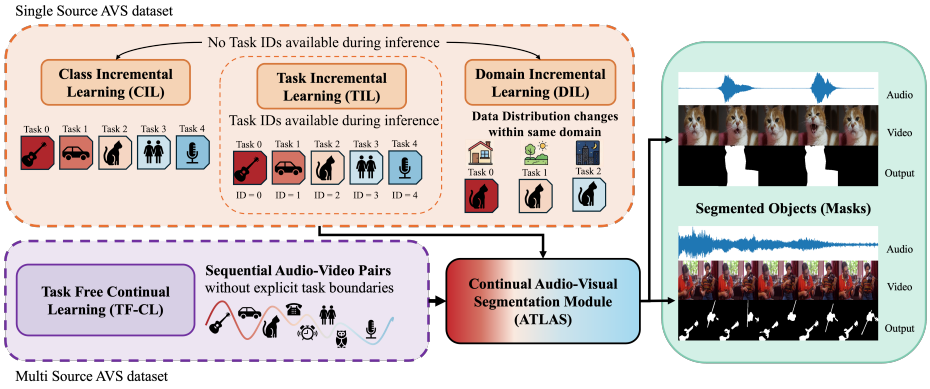


图 1： 音视频分割无样本持续学习基准概览

不会让听者忘记狗吠声。由于音频和视觉模态之间错综复杂的对应关系及相互依赖性、精确定位监督数据的稀缺，以及灾难性遗忘这一长期存在的问题 [20, 21, 27]，机器要复现这种终身学习能力仍然极具挑战性。

自提出以来，音视频分割已引起广泛关注，相关方法通过基于 Transformer 的融合 [50]、双向跨模态注意力 [15],以及时序音视频像素交互策略 [51]探索音视频对应关系。尽管取得了这些进展，一个关键方向仍未得到充分探索，因为现实世界中的音视频感知本质上是持续的。在实践中，已部署的模型会随时间遇到新的声音类别，范围从乐器和动物到车辆和人类语音。模型必须在不从头重新训练、也不回访先前数据的情况下学习这些类别。无样本持续学习（EFCL）为这一设定提供了原则性框架，使模型能够顺序获取新知识，同时在不存储任何数据的情况下缓解灾难性遗忘 [20, 21, 27]。预训练模型（PTMs）的出现进一步加速了这一方向的进展，因为建立在强大预训练表征之上的无样本持续学习方法 [14, 16, 38, 43, 49] 相比从头训练的框架展现出显著更好的现实适用性。尽管持续学习（CL）最初是在图像分类 [1, 22, 28, 36, 37, 48], 的背景下研究的，该领域此后已大幅扩展，包括持续语义分割 [47]、多模态持续学习 [46], 和持续强化学习 [35]，反映出向更现实、更开放式学习场景的更广泛转变。

然而，将这些方法直接应用于无样本持续学习设定，或将现有方法重新用于无样本持续学习设定，并非易事。音视频分割要求模型同时保持音频与视-

觉流之间的跨模态对齐，保留细粒度的空间分割边界，并维持先前学到的发声物体关联，且全程无需回顾或存储过去的数据。音视频分割的多模态特性会放大遗忘。任一模态的退化，或它们之间跨模态对齐的退化，都可能导致失败，即使每个模态仍保留有用信息。除遗忘之外，持续音视频分割还面临若干额外挑战。首先，新发声类别的引入会造成跨模态干扰，此时音频编码器与视觉编码器的特征空间都必须发生偏移以容纳新模式，从而可能使先前高度相关的模态之间出现错位。其次，评估复杂度 会增加，因为持续音视频分割需要密集的像素级指标，以便在兼顾先前遇到的各任务上的持续学习的同时，评估分割质量与音视频对齐。

为弥合这一差距，我们做出了三项贡献。**首先**，我们提出了音视频分割无样本持续学习基准（CL- AVS），如图 1 所示。该基准在两个音视频分割数据集上评估四种持续学习协议：在单源音视分割（SS- AVS）数据集 [51], 上评估任务增量学习（TIL）、类别增量学习（CIL）和领域增量学习（DIL），并在多源音视频分割（MS-AVS）数据集 [51]上评估无任务持续学习 [2] 。**其次**，我们提出了$ATLAS$，这是一种基于 LoRA [18]适配器构建的无样本基线，旨在实现参数高效持续学习。它在音视频融合前引入音频引导的融合前条件化，并采用低秩锚定（LRA）机制来减轻不同任务间的参数漂移。**第**三，我们针对多种持续学习算法和音视频分割框架开展了广泛实验，并通过详细分析揭示了持续音视频分割面临的挑战以及我们基线方法的有效性

2 相关工作

2.1 音视频分割

音视频分割（AVS）旨在通过联合建模音频与视觉信号，以像素级掩码定位视频帧中的发声物体。该任务在 AVSBench [51], 中提出，建立了二值音视频分割基准，随后扩展为具有类别感知预测的语义音视频分割 [50]。早期方法通常采用编码器–解码器架构，利用注意力或相关性机制将视觉表征与音频嵌入相融合。尽管这些模型在受控环境中有效，但在包含多个声源、背景噪声以及视觉显著但无声物体的真实场景中往往表现不佳 [8]。

近期音视频分割研究大体遵循两种设计范式。**视觉锚定**方法利用强大的视觉先验，并将音频视为条件信号，通常采用基于 Transformer 的融合或视觉基础模型来提升空间精度和边界质量 [19, 26, 33]。尽管效果显著，但过度依赖视觉线索可能会在音视频对应关系减弱时导致分割错误。相比之下，**音频锚定** 方法通过双向交互 [15] 或音频——

条件注意力机制 [6],来强调声音驱动定位，在视觉运动含糊不清时提升鲁棒性，但也会对噪声音频更加敏感。为降低标注成本，一些研究还通过利用自然的音视频对应关系或生成伪标签 [29,32]，进一步探索弱监督和自监督学习。

尽管音视频分割方法发展迅速，现有方法仍假设训练分布静态不变，所有类别都能同时获得。然而，在现实部署中，音视频系统会遇到不断变化的环境，新的发声类别也会随时间陆续出现。为弥合这一差距并奠定基础，我们提出了一个面向音视频分割的无样本持续学习基准。

2.2 持续学习

持续学习（CL）研究如何从顺序到达的数据中学习，同时缓解灾难性遗忘 [20,21,27]。早期研究主要关注图像分类，探索了基于回放、基于正则化和参数隔离等策略 [1,22,28,48]。随着大型预训练模型的兴起以及隐私限制日益严格，近期研究转向基于预训练表征的无回放适配 [14,16,43,49],，从而无需存储过去的数据即可实现可扩展的持续学习。

除分类之外，持续学习还拓展至持续语义分割等密集预测问题 [5,7,9,45,47]，以及多模态学习场景 [38,46]。将持续学习扩展到音视频分割还带来了额外挑战：模型在进行细粒度空间预测的同时，必须维持类别级知识和跨模态对齐。

3 方法

本节介绍数据集、持续音视频分割的问题定义，阐述定义持续学习音视频分割的基准协议，并最后介绍我们提出的 ATLAS 方法，以应对这一挑战。

3.1 数据集

为提供正式问题陈述所需的背景信息，我们首先介绍本研究中使用的数据集。我们采用了 AVSBench [51], 提出的数据集，具体包括单源音视（SS-AVS）和多源音视频分割（MS-AVS）数据集。SS-AVS 共包含 4,932 个视频，分为训练集、测试集和验证集，涵盖 23 个类别。该数据集采用半监督设置，每个视频仅首帧具有真实像素级分割标注。相比之下，MS-AVS 数据集包含 424 个视频，涵盖相同的 23 个类别。在这种设置下，视频按视频 ID 而非类别标签编索引，因为每个视频片段可能包含多个发声物体。MS-AVS 采用全监督设置，每一帧都有像素级标注，

但不提供类别标签。这两个数据集之间的这一根本差异，直接促使我们设计持续音视频分割（CL-AVS）基准协议。

3.2 问题形式化

音视频分割（AVS）的目标是学习一个模型，使其能够根据相应的音频信号，将视频帧划分为一组区域，从而定位发声物体。这些任务的区别在于预测掩码的具体形式。二元音视分割生成一组前景掩码，将发声物体与背景分隔开；语义音视分割则会为每个发声区域额外分配一个预定义物体类别集合中的类别标签。现在，我们对音视频分割中的持续学习（CL-AVS）进行一般化表述，以统一这两种设置，并将持续分割框架 [4, 11] 扩展到音视领域。

CL-AVS 在一系列 N 任务上训练模型（T_1，……，T_N），每个任务都会引入一组新的发声物体类别。在每个学习步骤 k，从数据集 $\mathcal{D}$ 中均匀抽取一定数量的 M 音视频对，以构建任务 $T_k = \{(V_{(k)i}, A_i^{(k)}, Y_{(k)i})\}_{i=1}^{M}$。这里，每个视频都表示为一系列帧 $V_i^{(k)} = \{v_{i,t}^{(k)}\}_{t=1}^{C_i}$，并配有时间上对齐的音频片段 $A_i^{(k)} = \{a_{i,t}^{(k)}\}_{t=1}^{C_i}$，其中 C_i 表示第 i 个视频的帧数。真实标注 $Y_i^{(k)} = \{y_{i,t}^{(k)}\}_{t=1}^{C_i}$ 因任务形式而异。令 $\mathcal{K}^k$ 表示在步骤 k 引入的发声物体类别集合，$\mathcal{K}^{1:k} = \bigcup_{j=1}^{k} \mathcal{K}^j$ 表示截至步骤 k 已见过的所有类别的累积集合，则，

$$y_{i,t}^{(k)} = \begin{cases} m_{i,t}^{(k)} \in \{0,1\}^{H \times W} & \textbf{binary setting} \\ s_{i,t}^{(k)} \in \mathcal{K}^{1:k} \cup \{0\}^{H \times W} & \textbf{semantic setting} \end{cases}$$

3.3 持续学习设置

持续学习主要有三种设置，即任务增量、类别增量和领域增量 [40]。我们将这些设置应用于基于单源音视分割（SS-AVS）数据集 [51] 的基准 CL-AVS，并进一步扩展出适用于多源音视频分割（MS-AVS）数据集 [51]的无任务 [23] 协议。

任务增量音视频分割：在此设置中，每个学习步骤 k 都会引入一组新的发声物体类别 $\mathcal{K}^k$。训练和测试时都会提供任务标识符（ID），以指明给定样本属于哪个步骤。测试时，模型会使用相应的任务标识符，在所有已见类别 $\mathcal{K}^{1:k}$ 上进行评估。

类别增量音视频分割：与任务增量设置类似，每个步骤 k都会引入新的发声物体类别 $\mathcal{K}^k$ 。然而，测试时不提供任务标识符。因此，模型必须在不知道给定样本来自哪个步骤的情况下，区分所有先前见过的类别 $\mathcal{K}^{1:k}$ 。

域增量音视分割：在该设置中，模型在单一发声物体类别（例如，吠叫的狗）内运行，并在同一类别但数据分布各异的视频上顺序学习，例如不同的视觉外观、场景或音频条件。形式上，类别集合在所有步骤中保持固定，即 $\mathcal{K}^1 = \mathcal{K}^2 = \cdots = \mathcal{K}^N$，而 $\mathcal{D}_k$ 的底层数据分布在每一步发生偏移。模型依据其在所有已遇到分布上保持分割性能且不遗忘先前分布的能力进行评估。

任务无关持续音视分割：我们将任务无关持续学习范式 [23]扩展到无法获得显式类别标签的多源音视分割数据集 [51], 。模型观察跨越 N 个边界模糊任务的视频流。目标是执行不进行语义区分的二值分割（发声与非发声），并在多个任务（例如， N=50）上依据随时推理原则 [23]进行评估。

3.4 ATLAS 框架

我们提出锚定稳定性的自适应任务学习（ATLAS），如图 2 所示，这是一种用于持续音视频分割的无样本基线。它能够在连续任务中从视频帧分割出发声物体，且无需重新训练或回访过去的数据。根据问题定义，对于给定任务 $T_k = \{(V_{(k)i}, A_i^{(k)}, Y_{(k)i})\}_{i=1}^{M}$，ATLAS 将每对 $(V_{(k)i}, A_i^{(k)})$ 映射为掩码逻辑值和类别逻辑值（可选），同时约束跨任务的参数漂移，以尽量减少持续学习过程中的灾难性遗忘。设视觉编码器和音频编码器分别为 $\Phi(\cdot;\Theta_v)$ 和 $\Phi(\cdot;\Theta_a)$。为实现参数高效适配，我们使用 LoRA [18]适配器训练视觉编码器骨干网络中选定的线性映射。对于每个预训练矩阵W_0，更新后的权重定义如下：

$$W = W_0 + \Delta W \text{ and } \Delta W = \frac{\alpha}{r} BA \tag{1}$$

其中 $A \in \mathbb{R}^{r \times d_{in}}$ 和 $B \in \mathbb{R}^{d_{out} \times r}$ 的秩为 $r \ll min(d_{in}, d_{out})$。超参数 r 用作常数缩放因子。LoRA 适配器通过更新视觉编码器的权重，有效地产生适配后的视觉特征。对于每个样本 i，视觉和音频编码后的特征为

$$X_{v,i}^{(k)} = \Phi_v(V_i^{(k)}; W) \text{ and } X_{a,i}^{(k)} = \Phi_a(A_i^{(k)}; \Theta_a) \tag{2}$$

这些视觉编码特征随后进入**音频引导的融合前条件化**模块，该模块将全局音频上下文注入视觉标记

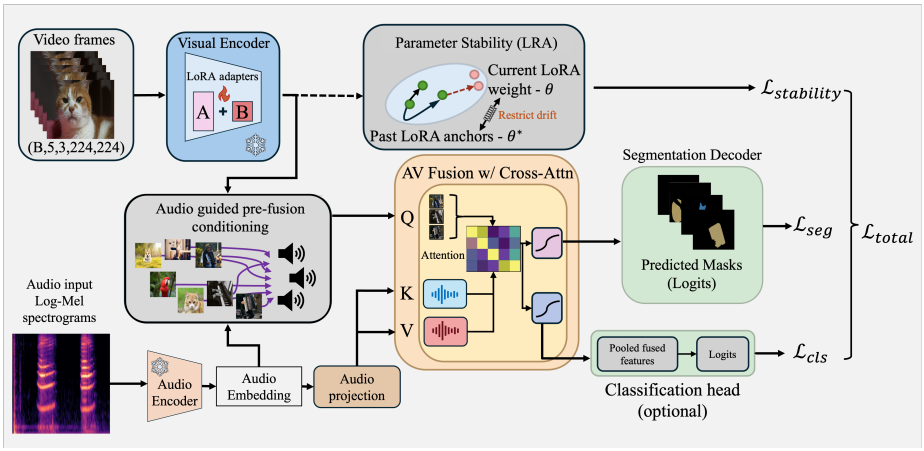


图 2：ATLAS 概览：该框架使用冻结的编码器骨干网络和视觉通路中的参数高效 LoRA 适配器，实现无需样本的持续音视频分割。音频引导的预融合条件化之后是交叉注意力，其中，经过条件化的视觉特征作为查询，音频特征作为键和值。为减轻跨任务的灾难性遗忘，我们的低秩锚定（LRA）模块会动态限制 LoRA 参数漂移。融合后的表示经解码后用于分割，并可选择使用分类分支。

通过学习得到的通道级调制 $\bar{X}_{v\,i}^{(k)}{}_{,\,i} = \mathcal{C}(X_{v\,i}^{(k)}{}_{,\,,}\ X_{a\,i}^{(k)}{}_{,\,i})$。音频引导的条件化算子为 $\mathcal{C}(\cdot)$（详细步骤见补充材料）。直观而言，这种学习得到的通道级调制可充当特征级门控机制。这里，音频特征被投影到视觉标记空间，以生成调制参数和条件化标记 $\bar{X}_{v\,i}^{(k)}{}_{,\,,}$。通过将音频上下文投影为缩放与平移参数，网络有选择地增强与发声物体相对应的特定视觉通道，同时抑制无关的背景噪声。在音视频融合之前，这一初步步骤会使视觉特征与声音相关区域对齐。随后，这些经过预条件化的视觉特征将作为后续**带交叉注意力的音视频融合**模块中的查询（Q）。我们分别根据这些经过预条件化的视觉特征和音频特征计算 Q、K 和 V：

$$Q_i^{(k)} = W_Q \bar{X}_{v,i}^{(k)}, \quad K_i^{(k)} = W_K X_{a,i}^{(k)}, \quad V_i^{(k)} = W_V X_{a,i}^{(k)}. \tag{3}$$

为在融合过程中保留视觉特征的空间完整性，融合特征通过对缩放点积注意力取 Softmax，并结合带层归一化 [41]的残差连接得到：

$$F_i^{(k)} = \texttt{LayerNorm}\left(\bar{X}_{v,i}^{(k)} + \texttt{softmax}\left(\frac{Q_i^{(k)} {K_i^{(k)}}^T}{\sqrt{d}}\right) V_i^{(k)}\right) \tag{4}$$

解码器 $\mathcal{D}$（其中也包含用于缓解灾难性遗忘的 LoRA 适配后的权重）将这些融合后的多模态特征投影至

原始视频帧的分辨率，以输出掩码逻辑值：

$$\hat{M}_i^{(k)} = \mathcal{D}\left(F_i^{(k)}\right) \tag{5}$$

此外，为支持类别 logits 的预测，全局池化层后接线性分类头 $\mathcal{H}$ ，用于处理组合表示并输出类别 logits

$$\hat{z}_i^{(k)} = \mathcal{H}\left(F_i^{(k)}\right) \tag{6}$$

对于我们的 CL-AVS 基准中的二元设置 $\left(y_{i,t}^{(k)} = m_{i,t}^{(k)}\right)$，分割采用二元交叉熵和 Dice [39] 损失的组合进行优化，以监督方式处理严重的前景与背景不平衡：

$$\mathcal{L}_{seg}^{(k)} = \frac{1}{M}\sum_{i=1}^{M}\left[\mathcal{L}_{BCE}\left(\hat{M}_i^{(k)}, m_i^{(k)}\right) + \mathcal{L}_{Dice}\left(\hat{M}_i^{(k)}, m_i^{(k)}\right)\right] \tag{7}$$

对于语义设置，我们以监督方式对迄今为止见过的类别预测 logits $Q^{1:k}$，分类项采用标准交叉熵（CE）：

$$\mathcal{L}_{cls}^{(k)} = \frac{1}{M}\sum_{i=1}^{M}\mathtt{CE}\left(\hat{z}_i^k, c_i^{(k)}\right) \tag{8}$$

其中， $c_i^{(k)} \in Q^{1:k}$ 是样本级类别标签。最后，我们引入 **低秩锚定（LRA）**模块，以尽量减小过去的锚定权重与当前权重之间的偏移，从而缓解灾难性遗忘。设 θ 表示当前可训练参数， θ^* 表示前一任务的锚定权重， Ω_i 表示参数重要性权重。我们不计算静态的 Fisher 近似，而是在训练过程中动态累积参数梯度与其更新量的乘积来计算 Ω_i 。这能有效追踪优化轨迹上的损失敏感度。稳定性正则化应用于 LoRA 矩阵和解码器权重。稳定性项定义如下：

$$\mathcal{L}_{stab} = \frac{c}{2}\sum_i \Omega_i\left(\theta_i - \theta_i^*\right)^2 \tag{9}$$

其中 c 是控制正则化强度的超参数。我们在任意任务 k 上的最终目标是公式7、8、9中所定义损失的组合

$$\mathcal{L}_{total}^{(k)} = \mathcal{L}_{seg}^{(k)} + \lambda_{cls}\mathcal{L}_{cls}^{(k)} + \mathcal{L}_{stab}^{(k)} \tag{10}$$

4 实验设置

4.1 实现细节

为确保评估严谨，我们选取具有代表性的音视频分割与持续学习方法，优先考虑公开可用的代码以及报告的较强

性能。对于非公开实现（以 *标记），我们使用大语言模型辅助开发（Anthropic Claude Code [3] 和Google Gemini [12]）重构了模型。

我们的实验设置涵盖四类密切相关、可适配于端到端 CL-AVS 训练的方法：(1) ( ) 通过将分类头替换（或添加）为分割头并配合音频编码器，从而适配于音视频分割的无样本持续学习方法；(2) ( ) 在持续数据流上训练的静态 AVS 模型；(3) ( ) 扩展了音频编码器与基于均方误差的跨模态对齐的持续语义分割框架；以及 (4) ( ) 基于回放的持续学习音频-视觉分割基线 [17]。

在所有实验中，我们使用 PANN [24]（在 AudioSet [13] 上训练）作为音频编码器。视觉骨干网络遵循其原始设计：包括 ATLAS 在内的大多数方法使用预训练的 ViT-B/16 [10], ，而 AVSBench、AVSBidirectional 以及 AVSegFormer、AVS-VCT [19] 保留 PVT-v2 骨干网络 [42]。对于 ATLAS，我们设置 $r = 8$、 $\alpha = 16$、 $\lambda_{cls} = 0.5$、 $c = 0.3$。我们使用 AdamW 对所有方法训练 30 个轮次，学习率为 10^{-3}，权重衰减为 5×10^{-4}。SS-AVS 采用 11-2 划分（共 7 个任务），MS-AVS 采用 31-5 划分（50 个任务）。结果在四块 NVIDIA A40 GPU 上以不同随机种子运行五次后取平均。

4.2 评估协议与指标

我们遵循标准的持续学习评估协议和持续语义分割指标 [30, 40, 47]。验证集仅用于模型选择，而留出的测试集则用于计算准确率矩阵**A** ，所有持续学习指标均由该矩阵衡量。在任务 t 上训练后，模型会在所有已见任务 ($k \leq t$) 以及一个未见任务（如有）上进行评估，以衡量前向迁移。在任务增量学习 (TIL) 中，评估时会提供任务标识；而类别增量、领域增量和无任务设置则要求在不提供任务信息的情况下进行推理。

1. **分割指标：**
   (a) **平均精度均值 (mAP)**衡量置信度分数的排序质量，不受固定阈值影响（↑越高越好）。

$$\mathrm{AP} = \sum_{j} \mathrm{Prec}(j), \Delta\mathrm{Rec}(j), \qquad \mathrm{mAP} = \frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} \mathrm{AP}_i. \tag{11}$$

   (b) **最大 F 分数 (Max-F)** 衡量不同决策阈值下可达到的最佳二元分割性能（↑越高越好）。

$$F_1(\tau) = \frac{2P(\tau)R(\tau)}{P(\tau) + R(\tau)}, \qquad \text{Max-F} = \max_{\tau \in [0,1]} F_1(\tau). \tag{12}$$

2. **持续学习指标：**

准确率矩阵 A 令 T 表示增量任务的总数。在学习任务 t 后，对任务 k 进行评估，得到

$$\mathbf{A} \in \mathbb{R}^{T \times T}, \qquad A_{t,k} = \text{performance after task } t \text{ evaluated on task } k. \tag{13}$$

该矩阵总结了整个任务序列中的性能表现。
(a) **学习准确率 (LA)** 每个任务首次学习时的性能（↑ 越好）。

$$\mathrm{LA} = \frac{1}{T} \sum_{t=0}^{T-1} A_{t,t}. \tag{14}$$

(b) **平均准确率 (AA)** 完成训练后在所有任务上的最终性能（↑ 越好）。

$$\mathrm{AA} = \frac{1}{T} \sum_{k=0}^{T-1} A_{T-1,k}. \tag{15}$$

(c) **遗忘 (F)**衡量学习新任务后先前任务上的性能下降（↓越好）。

$$\mathrm{F} = \frac{1}{T-1} \sum_{k=0}^{T-2} \left(\max_{t \in k,\ldots,T-1} A_{t,k} - A_{T-1,k} \right). \tag{16}$$

(d) **后向迁移 (BWT)**学习新任务对先前已学习任务的影响（正值 ↑ 为理想）。

$$\mathrm{BWT} = \frac{1}{T-1} \sum_{k=0}^{T-2} \left(A_{T-1,k} - A_{k,k} \right). \tag{17}$$

(e) **前向迁移 (FWT)**训练前对未来任务的零样本泛化（正值 ↑为理想结果）。

$$\mathrm{FWT} = \frac{1}{T-1} \sum_{k=1}^{T-1} A_{k-1,k}. \tag{18}$$

补充材料中还报告了其他分割指标（AUPR、mIoU 和 Dice）。

5 结果与讨论

表 1 展示了 SS-AVS 基准（任务增量学习、类别增量学习、域增量学习；11-2 划分，7 个任务）和 MS-AVS 基准（TF-CL；31-5 划分，50 个任务）上的结果。ATLAS 在全部四种设置中均取得最高平均精度均值，领先第二名 7 至 17 个百分点。在无样本增量学习方法中（），通过添加分割解码器和 PANN [24] 音频编码器来适配音视频分割的基于正则化

		SS-AVS（11-2 划分）						MS-AVS（31-5 划分）	
		TIL		CIL		DIL		TF-CL	
		平均 mAP	平均 For	平均 mAP	平均 For	平均 mAP	平均 For	平均 mAP	平均 For
	微调	27.91	7.85	28.04	10.89	25.38	9.30	16.61	11.58
	LwF [5] [IEEE Transactions on Pattern Analysis and Machine Intelligence '16]	26.70	10.34	26.89	7.52	24.79	12.52	21.59	13.43
	EWC [22] [美国国家科学院院刊 '17]	54.90	23.12	55.03	23.52	41.37	23.47	18.89	31.81
	突触智能 [48] [ICML '17]	59.10	13.74	56.52	15.57	42.44	22.77	19.51	33.25
	MAS [1] [欧洲计算机视觉会议 '18]	56.92	20.86	56.39	20.23	42.97	25.37	26.25	26.09
	L2P [43] [CVPR '21]	28.80	10.59	25.21	12.97	23.15	9.73	20.76	15.87
	RanPAC [49] [NIPS '23]	29.36	0.00	27.80	0.00	43.99	0.00	27.62	0.00
	FeCAM [14] [NIPS '23]	28.49	0.00	28.07	0.00	44.17	0.00	27.33	0.00
	DGR [16] [CVPR '24]	31.32	40.59	28.77	34.96	45.92	13.09	36.95	19.61
	PANDA [38] [人工智能促进协会 '26]	33.13	35.34	31.15	27.85	46.12	12.44	38.17	18.03
	AVSBench [50] [欧洲计算机视觉会议 '22]	63.84	22.05	62.28	23.64	52.78	24.95	31.15	21.09
	AVS-Bidirection [15] [人工智能促进协会 '24]	50.27	26.95	50.49	26.19	43.20	24.12	31.69	26.32
	COMBO [44] [CVPR '24]	27.67	7.38	27.18	8.89	23.07	9.95	23.76	11.02
	AVS-VCT [19] [CVPR '25]	25.64	8.60	25.54	11.18	22.19	13.49	21.98	13.87
	BalConpas [5] [欧洲计算机视觉会议 '24]	23.71	10.84	22.90	10.07	21.68	11.44	20.16	16.92
	EIR [45] [CVPR '25]	29.54	9.12	28.91	10.44	26.70	9.73	22.77	10.71
	CSTA* [7] [TCSV '25]	20.68	1.31	20.41	0.24	19.32	0.16	14.50	0.60
	HVPL* [9] [ICCV '25]	28.31	0.98	32.68	19.04	35.97	1.56	26.50	0.26
	CMR* [17] [arXiV '25]	22.75	0.32	19.16	2.77	19.59	1.30	25.92	1.35
	ATLAS（我们的方法）（我们的方法）	**74.67**	9.42	**71.39**	10.14	**63.84**	13.96	**45.27**	23.71

表 1：持续 AVS 基准上的性能 适配于 AVS 的无样本增量学习方法；扩展至持续学习的静态 AVS 模型；适配于 AVS 的持续分割；带回放的 CL-AVS 基线；以及我们的方法。性能通过各任务的平均精度均值（mAP）和平均反向遗忘度来衡量。

EWC [22], 突触智能 [48], 和 MAS [1]表现最佳，它们分别通过费舍尔信息、基于轨迹的累积以及输出敏感度，对任务重要权重的变化施加惩罚。在包含 50 个任务的 MS-AVS 上，它们逐权重的标量重要性却会导致性能下降，因为累积惩罚与大量任务和多个发声物体相冲突。L2P [43]通过提示池引导冻结的 ViT，但其提示仅作用于视觉标记，音频分支无法影响提示所作用的表征。RanPAC [49] 和 FeCAM [14] 通过完全冻结特征实现了零遗忘，但其平均精度均值仍接近微调基线。这表明，在没有自适应跨模态对齐的情况下，冻结视觉特征无法帮助定位发声物体。CMR [17] ( )会回放存储的样本，但仅取得 22.75（任务增量学习）平均精度均值，这说明若没有明确的视听定位，朴素回放并不足够。

在持续训练的静态 AVS 模型（）中，AVSBench [50] 凭借专门设计的视听交互模块，在 PVT-v2 上表现最佳，但缺乏持续学习机制，并且由于持续微调会覆盖已学习的融合权重，因此遗忘严重。AVS-Bidirection [15] 由于采用双向注意力，表现较差；参数更新会同时沿视觉和音频方向进行，从而加剧遗忘。COMBO [44] 和 AVS-VCT [19] 的表现接近微调基线，这表明其针对单任务设计的改进并未带来持续学习方面的提升。采用音频编码器并通过均方误差进行对齐的持续分割方法（）呈现出一种规律。HVPL [9] 和

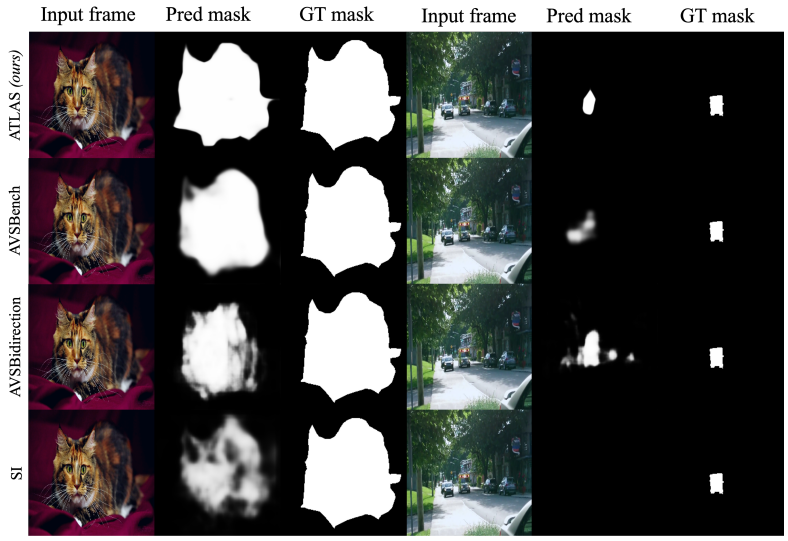


图3：在我们的 CL-AVS 设置中，排名前四的方法取得的 AVS定性结果：输入帧、中间帧的预测二值掩码以及真实掩码。（左）SS-AVS CIL 和 （右）MS-AVS TF-CL

CSTA [7] 的遗忘率极低，但平均精度均值表现一般，因为其单模态设计只更新少量参数，难以学习 AVS 所需的跨模态对应关系。DGR [16] 和 PANDA [38]通过梯度重加权和图块分布机制取得了良好表现，但较高的遗忘率也暴露了它们在适应两种模态方面的不足。我们将所提出模型 ATLAS 的性能提升归因于其主要组件。首先，LoRA 适配器将更新限制在视觉编码器和解码器的低秩子空间内。其次，音频引导的预融合条件化使视觉特征在执行交叉注意力之前，便朝着与声音相关的区域偏移。最后，我们的低秩锚定机制会动态惩罚适配器和解码器权重相对于先前任务锚点的漂移。此外，图3展示了表现最佳的四种方法的定性分割结果，凸显了我们方法的优势。

图4展示了在 SS-AVS 数据集采用 11-2 划分时，任务增量学习和域增量学习的前向迁移与遗忘情况。ATLAS 的前向迁移最高，且遗忘程度适中。大多数方法的前向迁移（FWT）都较低（< 0.45），要么遗忘严重（DGR、PANDA、MAS），要么以牺牲可塑性为代价保留知识（CSTA、FeCAM、RanPAC）。在域增量学习下，差距十分显著。ATLAS 的 FWT 超过 0.55，同时遗忘率与表现较差的模型相当。

图5展示了 SS-AVS 中类别增量学习设置下的最大F值热图。ATLAS 在各任务中都保持着明亮的单元格，甚至到第 6 行时，早期任务的表现依然出色，这表明我们的低秩锚定机制能够有效限制先前学得表征的漂移。AVSBench 在早期任务上表现良好，但

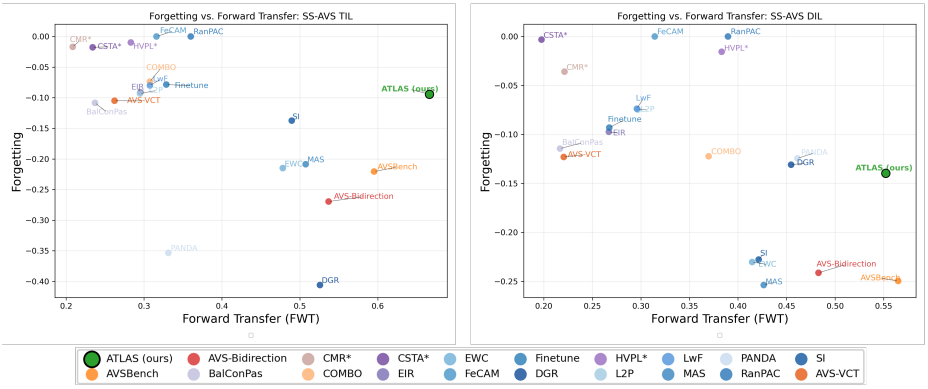


图4：前向迁移与遗忘：CL-AVS 基准中各模型在SS-AVS数据集上、于任务增量学习与域增量学习协议下前向迁移与遗忘之间的权衡。

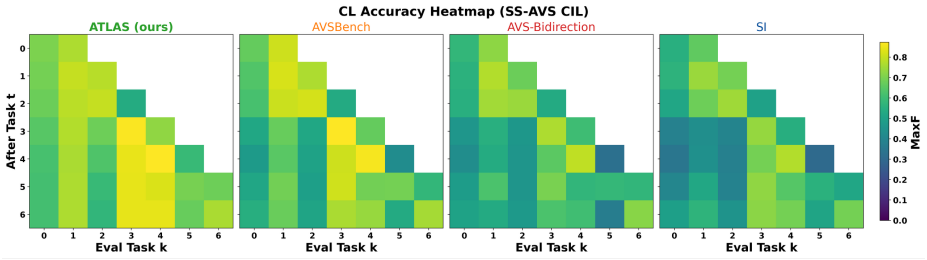


图5：MaxF分数热力图在采用11–2划分的CIL协议下，于SS-AVS测试集上在每个增量任务之后进行评估。为清晰起见，纳入了四种表现最佳的方法。

随着新任务的引入，其表现逐渐下降，因为它并非为持续学习而设计。AVS-Bidirectional 在任务 3 之后表现骤降，这可能是由于错误在其双向交互中传播所致。相比之下，突触智能能够保留各参数的重要性，但难以维持准确分割所需的分布式跨模态表征。补充材料中提供了更多结果和可视化内容，包括超参数分析、其他指标，以及采用不同骨干架构的实验。我们注意到，持续学习器的性能在很大程度上取决于预训练模型的质量。

6 消融研究

在本节中，我们通过消融研究分析方法的效率–性能权衡，以及各个组成部分的贡献。图 6 展示了六项指标的归一化雷达图轮廓：准确率、保持性、

			SS-AVS（11-2 划分）				MS-AVS（31-5 划分）			
方法	预条件	LRA	平均 mAP	平均 For	平均 F1	平均 mIoU	平均 mAP	平均 For	平均 F1	平均 mIoU
ATLAS	✓	✓	**74.67**	**9.42**	**62.76**	**50.85**	**47.53**	23.71	**42.27**	**39.44**
ATLAS	✗	✓	71.17	12.11	59.97	47.78	43.12	25.07	39.38	38.2 7
ATLAS	✓	✗	67.18	18.03	50.38	39.16	34.41	26.61	36.72	24.12
ATLAS	✗	✗	65.33	21.00	47.31	33.58	29.19	27.01	31.16	20.12
AVSBench	-	-	63.84	22.05	50.63	37.28	31.15	**21.09**	34.39	20.15

表 2：ATLAS 的组件消融实验。我们分析了音频引导预条件和低秩锚定（LRA）对 CL-AVS 基准的贡献

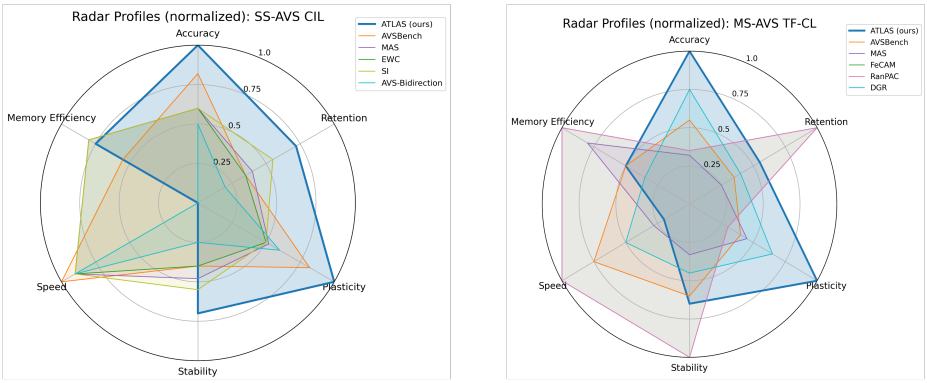


图6：雷达图 总结了在准确率、保持性、可塑性、稳定性、速度和内存效率这六个维度上的归一化性能，对象为在CIL协议下的SS-AVS数据集以及在TF-CL协议下的MS-AVS数据集上表现最佳的六种方法。

可塑性、稳定性、速度和内存效率。对于 SS-AVS（类别增量学习）和 MS-AVS（TF-CL），ATLAS 覆盖的面积均最大，尤其在准确率和可塑性方面表现出色。正则化方法的总体覆盖面积较小，而 RanPAC 和 Fe-CAM 等基于原型的方法虽然内存效率较高，但往往会欠拟合。为了解 ATLAS 各组成部分的作用，我们进行了消融实验，结果汇总于表 2，并以 AVSBench 作为基线进行比较。结果表明，低秩锚定（LRA）是最关键的组成部分，因为它能稳定 LoRA 适配后的权重，并减少持续学习过程中的参数漂移。音频引导的预融合条件处理带来了额外收益，但贡献不及 LRA。即使没有这些模块，基于 LoRA 的模型在不同任务间仍会发生参数漂移，导致其表现低于 AVSBench，这种情况在任务数量较多的 MS-AVS 上尤为明显。

7 结论

本文提出了一种面向音视频分割的无样本持续学习基准，旨在生成视频中发声物体的像素级掩码

。我们在单源和多源音频视觉分割数据集上，按照四种持续学习协议评估了多种方法。此外，我们提出了 ATLAS，这是一种端到端的无样本基线，能够在提升学习效果的同时减少表征漂移。实验结果证明了该方法的有效性，并为未来的持续音视学习研究奠定了基础。

参考文献

1. Aljundi, R., Babiloni, F., Elhoseiny, M., Rohrbach, M., Tuytelaars, T.：记忆感知突触：学习该记住什么（以及不该忘记什么）（2018），https://arxiv.org/abs/1711.09601 2. Aljundi, R., Kelchtermans, K., Tuytelaars, T.：无任务持续学习。载于：IEEE计算机视觉与模式识别会议（CVPR）（2019年6月）3. Anthropic：推出新一代Claude。https://www.anthropic. com/news/claude-3-family（2024）4. Cermelli, F., Cord, M., Douillard, A.：Comformer：语义分割与全景分割中的持续学习。载于：IEEE/CVF计算机视觉与模式识别会议（2023）5. Chen, J., Cong, R., Luo, Y., Ip, H.H.S., Kwong, S.：在持续全景分割中取得平衡。载于：欧洲计算机视觉会议（2024）6. Chen, T., Tan, Z., Chu, Q., Wu, Y., Liu, B., Yu, N.：通过强化音频线索引导音视频分割。载于：国际学习表征会议（ICLR）研讨会（2024），https://openreview.net/forum? id=WdWGe88RdX，通过促进音频与视觉特征之间更深入的双向交互来缓解模态不平衡——这是一种*音频锚定交互*视角。7. Chen, T., Liu, H., Lim, C.H., See, J., Gao, X., Hou, J., Lin, W.：Csta：用于无样本视频类别增量学习的时空因果自适应学习。IEEE视频技术电路与系统汇刊**35**(11), 11488–11501 (2025) 8. Chen, Y., Liu, Y., Wang, H., Liu, F., Wang, C., Frazer, H., Carneiro, G.：揭示实例关联：深入探究音视频分割。arXiv预印本 arXiv:2304.02970（2023），https://arxiv.org/abs/2304.02970，分析音视分割中的跨模态对齐与基准偏差 9. Dong, J., Yin, H., Liang, W., Zhao, H., Ding, H., Sebe, N., Khan, S., Khan, F.S.：用于持续视频实例分割的层次化视觉提示学习。载于：IEEE/CVF国际计算机视觉会议（ICCV）（2025年10月）10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.：一张图像胜过16×16个词：用于大规模图像识别的变换器。ICLR（2021）11. Douillard, A., Chen, Y., Dapogny, A., Cord, M.：PLOP：持续语义分割中的无遗忘学习。载于：IEEE/CVF计算机视觉与模式识别会议（CVPR）（2021）12. Gemini团队，Google DeepMind：Gemini：一系列能力强大的多模态模型。arXiv预印本 arXiv:2312.11805（2023）13. Gemmeke, J.F., Ellis, D.P.W., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.：Audio Set：用于音频事件的本体与人工标注数据集。载于：2017 IEEE国际声学、语音与信号处理会议（ICASSP）。第776–780页（2017）。https://doi.org/10.1109/ ICASSP.2017.7952261 14. Goswami, D., Liu, Y., Twardowski, B., van de Weijer, J.：FeCAM：利用类别分布异质性实现无样本持续学习。载于：第37届神经信息处理系统国际会议论文集。NIPS '23，Curran Associates Inc.，美国纽约州雷德胡克（2023）

15. Hao, D., Mao, Y., He, B., Han, X., Dai, Y., Zhong, Y.：通过双向生成改进音视频分割。载于：AAAI 人工智能会议论文集（2023 08288 16 2308），https://arxiv.org/abs/。He, J.：梯度重加权：迈向不平衡类别增量学习。IEEE/CVF计算机视觉与模式识别会议论文集（CVPR），第16668–16677页（2024年6月）17. Hong, Y., Yang, Q., Zhang, T., Wang, Z., Fu, Z., Ding, K., Fan, B., Xiang, S.：驾驭持续音视频分割中的模态纠缠（2025），https://arxiv.org/abs/2510 17234 18。Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.：LoRA：大语言模型的低秩适配（2021），https://arxiv. org/abs/2106 09685 19。Huang, S., Ling, R., Hui, T., Li, H., Zhou, X., Zhang, S., Liu, S., Hong, R., Wang, M.：使用视觉中心型Transformer重新审视音视频分割。载于：IEEE/CVF计算机视觉与模式识别会议（CVPR）（2025），https://openaccess.thecvf.com/content/CVPR2025/papers/Huang_Revisiting_Audio-Visual_Segmentation_with_Vision-Centric_Transformer_CVPR_2025_paper.pdf，提出一种*视觉中心型*变换器，由视觉查询引导音频融合，从而提升轮廓质量与分离准确度。20. Kemker, R., McClure, M., Abitino, A., Hayes, T., Kanan, C.：测量神经网络中的灾难性遗忘。《人工智能促进协会人工智能会议论文集》**32**(1)（2018年4月）21. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., Hadsell, R.：克服神经网络中的灾难性遗忘。《美国国家科学院院刊》**114**(13), 3521–3526（2017年3月）22. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., Hadsell, R.：克服神经网络中的灾难性遗忘。《美国国家科学院院刊》**114**(13), 3521–3526（2017）。https://doi.org/10.1073/pnas.1611835114，https://www.pnas.org/doi/abs/10 1611835114 23 1073。/pnas.。Koh, H., Kim, D., Ha, J.W., Choi, J.：在类别增量模糊任务配置下进行在线持续学习，并实现随时推理（2022），https://arxiv.org/abs/ 2110.10031 24. Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., Plumbley, M.D.：PANNs：用于音频模式识别的大规模预训练音频神经网络（2020），https: //arxiv.org/abs/1912 10211 25。Leavitt, V.M., Molholm, S., Gomez-Ramirez, M., Foxe, J.J.：“什么”与“哪里”：对支撑声音对象识别和声音定位的不同神经过程进行高密度电生理映射研究。《整合神经科学前沿》**5**, 23（2011年6月）26. Li, J., Zhao, W., Huang, Z., Guo, Y., Tian, Y.：音视频分割模型真的能分割发声物体吗？arXiv预印本 arXiv:2502.00358（2025），https://arxiv.org/abs/2502.00358，研究视觉主导偏差以及负向音频下的鲁棒性。27. Li, X., Zhou, Y., Wu, T., Socher, R., Xiong, C.：学会扩展：一种克服灾难性遗忘的持续结构学习框架。载于：Chaudhuri, K., Salakhutdinov, R.（编）《第36届机器学习国际会议论文集》。机器学习研究论文集，第97卷，第3925–3934页。PMLR（2019年6月9–15日），https://proceedings.mlr.press/v97/li19m.html

28. Li, Z., Hoiem, D.: 无遗忘学习。《IEEE Transactions on Pattern Analysis and Machine Intelligence》**40**, 2935–2947 (2016), https://api. semanticscholar.org/CorpusID:4853851 29. Liu, J., et al.: 无标注音频视觉分割。WACV (2024), https: //openaccess.thecvf.com/content/WACV2024/papers/Liu_ Annotation-Free_Audio-Visual_Segmentation_WACV_2024_paper.pdf，通过匹配图像掩码与音频，以最少的标注合成音频视觉分割数据集，并将 SAM 适配至音视频分割。30. Lopez-Paz, D., Ranzato, M.: 持续学习的梯度情景记忆 (2022 08840 31 1706), https://arxiv.org/abs/ . . Mihalik, A., Noppeney, U.: 视听感知中的因果推断。《神经科学杂志》**40**(34), 6600–6612 (2020). https://doi.org/10.1523/JNEUROSCI. 0051 6600 32 20 2020, 40 34 - . https://www.jneurosci.org/content/ / / . Mo, S., Raj, B.: 弱监督音频视觉分割。载于：NeurIPS 音视机器学习研讨会 (2023), https://openreview.net/forum? id=sUqG96QqZM，提出了采用多尺度对比学习的弱监督音视频分割方法。33. Mo, S., Tian, Y.: Av-sam：任意分割模型与音视频定位和分割相结合。arXiv 预印本 arXiv:2305.01836 (2023), https://arxiv. org/abs/2305.01836，探索通过逐像素融合将 SAM 与音视频分割相结合，代表了早期的*视觉锚定基础模型*方法。34. Newell, F.N., McKenna, E., Seveso, M.A., Devine, I., Alahmad, F., Hirst, R.J., O' Dowd, A.: 多感官知觉制约物体类别的形成：关于感官驱动和预测过程对类别决策影响的证据综述。Philos. Trans. R. Soc. Lond. B Biol. Sci.**378**(1886), 20220342 (2023年9月) 35. Pan, C., Yang, X., Li, Y., Wei, W., Li, T., An, B., Liang, J.: 持续强化学习综述 (2025 21872 36 2506), https://arxiv.org/abs/ . . Raghavan, S., He, J., Zhu, F.: Delta：解耦长尾在线持续学习。载于：IEEE/CVF计算机视觉与模式识别会议论文集 (2024) 37. Raghavan, S., He, J., Zhu, F.: 面向真实场景食品图像分类的在线类别增量学习。载于：IEEE/CVF 冬季计算机视觉应用会议 (WACV) 论文集。第 8195–8204 页 (2024年1月) 38. Raghavan, S., He, J., Zhu, F.: Panda——面向长尾无样本持续学习的补丁与分布感知增强。载于：AAAI 人工智能会议论文集 (2025 09791 39 2511), https://arxiv.org/abs/ . . Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Jorge Cardoso, M.: 作为高度不平衡分割深度学习损失函数的广义 Dice 重叠。Deep Learn. Med. Image Anal. Multimodal Learn. Clin. Decis. Support (2017)**2017**, 240–248 (2017年9月) 40. Sun, H.L., Zhou, D.W., Ye, H.J., Zhan, D.C.: Pilot：基于预训练模型的持续学习工具箱。arXiv 预印本 arXiv:2309.07117 (2023) 41. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: 注意力就是一切。载于：第31届神经信息处理系统国际会议论文集。第 6000–6010 页。NIPS'17, Curran Associates Inc., Red Hook, NY, USA (2017) 42. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: PVT v2：采用金字塔视觉变换器的改进基线。《计算视觉媒体》**8**(3), 415–424 (2022年9月)。https://doi.org/10.1007/s41095-022-0274-8, http://dx.doi.org/10.1007/s41095-022-0274-8

43. Wang, Z., Zhang, Z., Lee, C.Y., Zhang, H., Sun, R., Ren, X., Su, G., Perot, V., Dy, J., Pfister, T.: 面向持续学习的提示学习。载于：IEEE/CVF计算机视觉与模式识别会议论文集。第 139–149 页 (2022) 44. Yang, Q., Nie, X., Li, T., Gao, P., Guo, Y., Zhen, C., Yan, P., Xiang, S.: 协作确实重要：探索音视频分割中的多阶双边关系 (2024) 45. Yin, H., Feng, T., Lyu, F., Shang, F., Liu, H., Feng, W., Wan, L.: 超越背景偏移：重新思考持续语义分割中的实例回放 (2025 22136 46 2503), https://arxiv.org/abs/ . . Yu, D., Zhang, X., Chen, Y., Liu, A., Zhang, Y., Yu, P.S., King, I.: 多模态持续学习的最新进展：全面综述 (2024), https://arxiv. org/abs/2410 05352 47 . . Yuan, B., Zhao, D.: 持续语义分割综述：理论、挑战、方法与应用。《IEEE Transactions on Pattern Analysis and Machine Intelligence》**46**(12), 10891–10910 (2024年12月)。 https://doi.org/10.1109/ tpami.2024 3446949 48 3446949, 10 1109 2024 . http://dx.doi.org/ . /TPAMI. . . Zenke, F., Poole, B., Ganguli, S.: 通过突触智能实现持续学习。载于：第34届国际机器学习会议论文集——第70卷。第 3987–3995 页。ICML'17, JMLR.org (2017) 49. Zhou, D.W., Ye, H.J., Zhan, D.C., Liu, Z.: 借助预训练模型重新审视类别增量学习：泛化性和适应性就是你所需要的一切。《神经信息处理系统论文集》(2023) 50. Zhou, J., Shen, X., Wang, J., Zhang, J., Sun, W., Zhang, J., Birchfield, S., Guo, D., Kong, L., Wang, M., Zhong, Y.: 结合语义的音视频分割。《国际计算机视觉期刊》页码 1–21 (2024) 51. Zhou, J., Wang, J., Zhang, J., Sun, W., Zhang, J., Birchfield, S., Guo, D., Kong, L., Wang, M., Zhong, Y.: 音视频分割。载于：欧洲计算机视觉会议 (2022)