arXiv:2301.10295v1 [cs.CV] 4 Jan 2023

18-786 项目最终报告：结合音频上下文的物体分割

Kaihui Zheng
电气与计算机工程 卡内基梅隆大学
kaihuiz@andrew.cmu.edu

Yuqing Ren
电气与计算机工程 卡内基梅隆大学
yuqingr@andrew.cmu.edu

Zixin Shen
电气与计算机工程 卡内基梅隆大学
zixins@andrew.cmu.edu

Tianxu Qin
电气与计算机工程 卡内基梅隆大学
tianxuq@andrew.cmu.edu

摘要

在含音频的视频录制中，视觉对象通常具有与其自然同步的声学特征。本项目探索视频实例分割任务中的多模态特征聚合，将音频特征整合到视频分割模型中，以实现视听学习方案。我们的方法基于现有的视频实例分割方法，该方法利用视频帧之间丰富的上下文信息。由于这是首次研究视听实例分割，我们收集了一个新数据集，其中包含 20 个发声类别及其同步的视频和音频录制。通过使用组合解码器融合视频和音频特征，我们的模型相较于基础模型取得了小幅提升。此外，我们还通过广泛的消融实验展示了不同模块的有效性。

1 引言

受益于视频领域的发展，视频实例分割（VIS）应运而生并受到广泛关注。VIS 旨在对各帧中的目标掩码进行分割，同时保持其身份在整个视频中的一致性。以往的方法专注于通过单一视频模态提升分割质量，却忽视了相应音频录制中蕴含的丰富信息。本报告不仅将同时考虑视觉和听觉信号来研究视频实例分割任务，还将采用不同结构实现我们的模型。

对于单模态音频感知问题，以往的方法在多个任务上都取得了令人瞩目的表现，例如事件检测 [2], 语音识别 [6] 和声音分类[32]。对于单模态视频理解，分类 [3], 分割 [26] 和跟踪 [18] 等任务也取得了巨大成就。然而，尽管在现实世界应用中，视频和音频信号总是共存的，但以往的方法在处理时却将二者分开。

我们生活的世界中有大量的视觉和听觉信号。我们的系统能够识别物体和信号，接着分割物体覆盖的图像区域，并最终以联合方式分离这些物体发出的声音。尽管已有大量研究探讨同步的视觉和听觉信号之间的关联，但相关应用

深度学习导论 (11785/18786) 项目报告。

仅聚焦于若干粗粒度任务，例如事件检测 [36], 声音分离 [35]和目标定位 [27]。在许多情况下，音频的功能层面常被忽视。值得注意的是，听觉信号包含发声物体的语义和位置信息，以及背景环境的上下文信息。这些特征在视频领域的密集预测任务中具有很大的利用潜力。

在本项目中，我们成功展示了 CrossVis 带来的显著性能提升。在此基础上，我们尝试探索基于视频和音频数据的联合学习方案。具体而言，我们研究了视频实例分割任务中视频与音频之间的相关性。首先，我们根据视频帧率对音频频谱图数据进行采样。随后，我们将两种信号结合起来，并将融合后的数据分别送入不同的全卷积网络。经过分割模型处理后，我们根据分割结果衡量了性能提升。

CrossVis [31] 利用全卷积单阶段目标检测器（FCOS）中的实例特征，在其他帧中定位同一实例。

2 相关工作

2.1 视频实例分割

近年来，互联网上产生了海量的视频数据。手动从如此庞大的互联网视频数据中提取语义信息极不现实，因此需要自动化方法从视频数据中标注有用信息 [10]。因此，视频分割是视频处理与检索的重要步骤之一。视频实例分割[31, 9, 19, 5, 22, 23] 是在图像分割 [12] 基础上扩展而来的任务，并加入了实例身份跟踪。视频目标分割 [30, 17, 14, 33, 28] 旨在将视频中的每一帧划分为有意义的对象，即沿时空方向对外观和运动具有连贯性的像素进行分组 [11]。视频目标分割任务极具挑战性，原因如下：（i）视频中的对象数量未知；（ii）视频背景不断变化；（iii）视频中会出现多个对象 [1]。现有的视频分割方法大致可分为两类：交互式方法和无监督方法。交互式物体分割方法在初始化过程中需要人工干预，而无监督方法能够自动执行物体分割。在半监督方法 [4],中，用户需要标注初始帧，再将这些标注传播到视频的所有帧中。近年来，多模态视觉日益流行，它利用文本、音频或传感器数据 [15, 25, 21, 16] 来辅助视觉输入。一些最新研究聚焦于多模态视频目标分割，利用文本 [29], 、音频 [20, 24] 或无线信号 [34]。

2.2 视听学习

为了学习输入声音与图像之间的关系，并更好地预测它们的表征，冯子舜及其团队提出了一种方法，不仅在乐器数据上，还在野外视频上探索音视对应（AVC）性能。根据论文野外视频的自监督视听表征学习。[13] 作者使用给定视频数据集中的视频帧样本和 1 秒音频片段。模型的输入是一对单个视频帧和一个音频片段，并使用两个 ResNet-18 网络 [8] 分别提取视觉特征和音频特征。模型通过 Sigmoid函数计算视觉特征与音频特征之间的距离作为激活值，最终预测输入视频帧与音频片段是否对应。该模型以自监督方式进行训练。作者使用 VGGSound数据集[7]训练模型。训练参数为 400k 训练步数，批量大小为 256。训练总共使用了超过 1 亿对采样的音频-图像对。优化器为 Adam优化器，学习率为 0.001。结果表明，该方法可用于在各种场景下获得良好的检索结果，且与其他方法相比取得了有竞争力的结果。

3 方法

3.1 模型描述

我们在模型中整合了音频特征与视觉特征（图 1），采用的方法来自《快速在线视频实例分割的交叉学习》[31]，这是一个基于交叉学习的视频实例分割框架。在所有已发表的方法中，该框架达到了最先进的性能。我们基于交叉学习模型构建了自己的方法，并尝试在后续阶段加入音频线索加以改进。所提出的模型名为 CrossVis，由两个关键组件构成：用于更准确地学习视频实例表征的交叉学习方案，以及全局平衡实例嵌入分支。作者最初利用了不同视频帧间的丰富信息。首先，对于静态图像实例分割，我们使用动态条件卷积生成实例掩码$\mathbf{M}_{x,\ y}$˜方法是将掩码分支中的特征图$\mathbf{F}_{x,\ y}$与一组实例特定动态滤波器$\theta_{x,\ y}$进行卷积

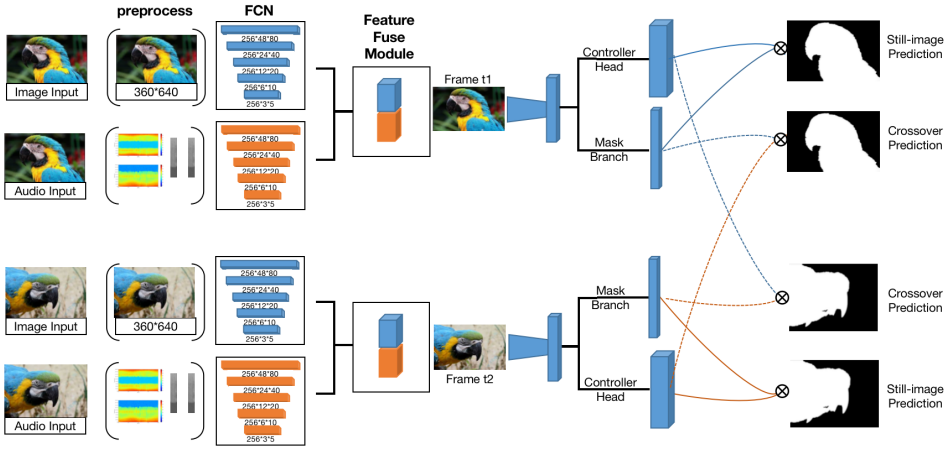

图 1：Audio Fusion CrossVis。我们的模型基于 CrossVis，将音频特征整合到视觉特征中。我们在 CrossVis 基础上采用的主要技巧是使用两个基线模型，分别从音频和视觉信号中提取特征。提取信号特征后，我们将这两个特征图拼接起来，并通过卷积将其融合为一个特征，然后进行 CrossVis 论文中提出的交叉学习。

$$\tilde{\mathbf{F}}_{x,y} = Concate(\mathbf{F}_{mask}; \mathbf{O}_{x,y})$$
$$\mathbf{M}_{x,y} = MaskHead(\tilde{\mathbf{F}}_{x,y}; \theta_{x,y})$$

其次，对于视频实例分割任务，给定从一个视频中采样的一对帧，同一实例可能出现在不同位置。因此，作者利用一帧的外观信息 $\theta_{x,\ y}$ 将另一采样帧中同一实例的位置信息 $O_{x,\ y}(t, \delta)$ 融合进来。在每一帧中，在时间 t，位于 $(x,\ y)$ 的 $I_i(t)$ 的实例掩码可表示为：

$$\mathbf{M}_{x,y}(t) = MaskHead(\tilde{\mathbf{F}}_{x,y}(t); \theta_{x,y}(t))$$

在时间 $t+\delta$，该实例从位置 $(x,\ y)$ 移动到位置 $(x',\ y')$。因此，$I_i(t+\delta)$ 的实例掩码可表示为：

$$\mathbf{M}_{x',y'}(t+\delta) = MaskHead(\tilde{\mathbf{F}}_{x',y'}(t+\delta); \theta_{x',y'}(t+\delta))$$

除此之外，该模型巧妙的交叉学习方案建立了某一帧的动态滤波器与另一帧的掩码特征图之间的联系。我们期望动态滤波器 $\theta_{x,y}(t)$能够通过卷积其掩码特征图˜$\mathbf{F}_{x',y'}(t+\delta)$，生成另一时刻帧的掩码：

$$\mathbf{M}^*_{x',y'}(t+\delta) = MaskHead(\tilde{\mathbf{F}}_{x',y'}(t+\delta); \theta_{x',y'}(t))$$

与时间 t 的掩码相同：

$$\mathbf{M}^*_{x,y}(t) = MaskHead(\tilde{\mathbf{F}}_{x,y}(t); \theta_{x',y'}(t+\delta))$$

其中 $\mathbf{M}^*$ 带有上标“*”，表示通过交叉学习生成的实例掩码。根据以上四个公式，在训练过程中，我们均通过 Dice 损失进行优化：

$$L_{dice}(\mathbf{M}, \mathbf{M}*) = 1 - \frac{2\sum_i^{HW} \mathbf{M}_i \mathbf{M}_i^*}{\sum_i^{HW} (\mathbf{M}_i)^2 + \sum_i^{HW} (\mathbf{M}_i^*)^2}$$

4 实验

4.1 数据集

我们收集了一个包含视频及其对应音频录制的数据集。具体而言，该数据集由两部分组成：Youtube-VIS-2019（人工标注）和自采集数据集（自动标注）。Youtube-VIS-2019 是一个基于初始 Youtube-VOS 数据集的视频实例分割数据集，其类别标签集包含人、动物和车辆等 40 种常见物体，共有 4883 个独立的视频实例和 131k 条高质量人工标注。Youtube-VIS-2019 包含 2883 个高分辨率 YouTube 视频，划分为 2338 个训练视频、302 个验证视频和 343 个测试视频。图 2 展示了我们数据集的一个示例。

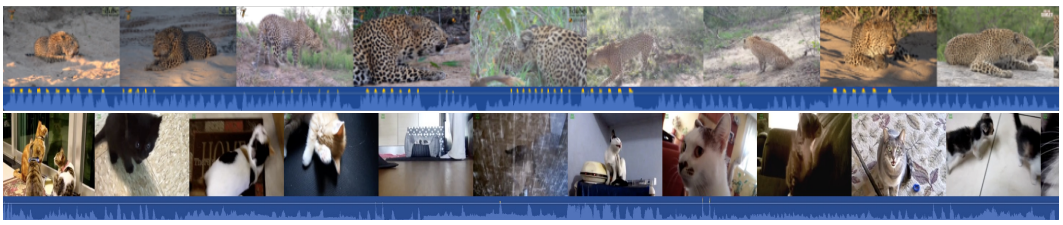

图 2：数据集。第一行和第二行分别是豹类和猫类。

4.2 数据预处理

为了从视觉和音频信号中提取更好的特征并获得更好的实验结果，我们在将视觉和音频信号输入模型之前对其进行了预处理。

4.2.1 视频

首先，我们以特定帧率从视频中提取图像帧。然后将所有图像调整为相同大小。最后，我们使用目标检测框架 Detectron2 中的 ResizeShortestEdge 函数进行数据增强。

4.2.2 音频

我们首先将双声道音频转换为单声道音频，以简化后续计算。图 4 中的频谱是信号频率谱随时间变化的可视化表示。我们将单声道音频预处理为频谱，并移除没有对应视频的音频。然后，根据视频的帧率，沿表示时间的 X 轴均匀划分频谱。以下图片展示了双声道/单声道音频频谱图和采样音频频谱图。图 3

4.3 基线选择

CondInst 实例分割模型遵循 Detectron2 和 AdelaiDet 的 1 * 训练计划，在 COCO train2017 上进行预训练。我们使用该模型初始化 CrossVIS。随后，我们在 Youtube-VIS-2019 数据集上按照 1 * 训练计划训练 CrossVIS，即训练 12 轮次。根据 SipMask-VIS，将初始学习率设为 0.005。在第 9 轮和第 11 轮，学习率分别降低 10 倍。单尺度训练时，帧大小为 360*640。多尺度训练时，帧大小与 SipMask-VIS 中的设置相同。推理时，帧大小为 360 * 640。我们在 YouTube-VIS-2019 上评估 CrossVIS。除非另有说明，AP 指精确率-召回率（PR）曲线下的面积。AP 是在多个交并比（IoU）阈值下取平均值。

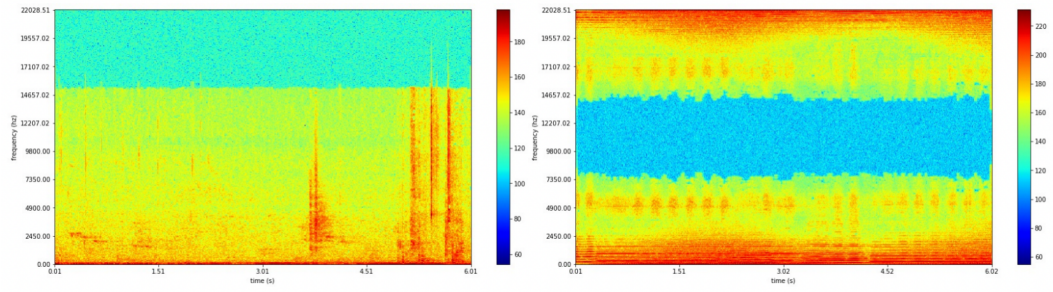

图 3：双声道/单声道音频频谱图。

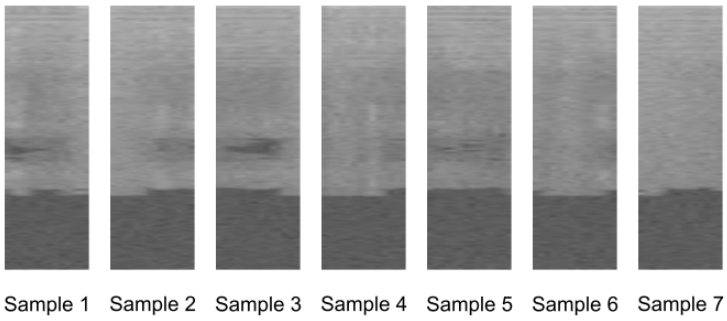

图 4：采样音频频谱图。

本报告中的 AR 指平均召回率，其定义为每个视频给定固定数量的分割实例时所能达到的最大召回率。与以往工作一致，我们报告在验证集上的结果，以评估所提方法的有效性。

4.4 主要结果

4.4.1 基线模型实现

首先，我们实现了 CrossVis 模型以获得分割结果。经过 22559 次迭代后，如图 5 所示，总损失降至 1.132，其中 focos 分类损失为 0.08897，focos 定位损失为 0.008727，掩码损失为 0.06239，交叉学习损失为 0.06859，嵌入损失为 0.2318。

实现后，我们尝试调整模型的最佳参数。例如，我们使用 WarmupMultiStepLR 作为学习率调度器来训练模型，这意味着在最初几个轮次中使用较小的学习率。随着模型趋于稳定，我们会提高学习率进行训练。一个轮次内的学习率变化如图 6 所示。

4.4.2 消融研究

从 CrossVis 的论文中，我们了解到研究人员尝试采用一种新颖的交叉学习方案，利用当前帧中的实例特征，逐像素定位其他帧中的同一实例

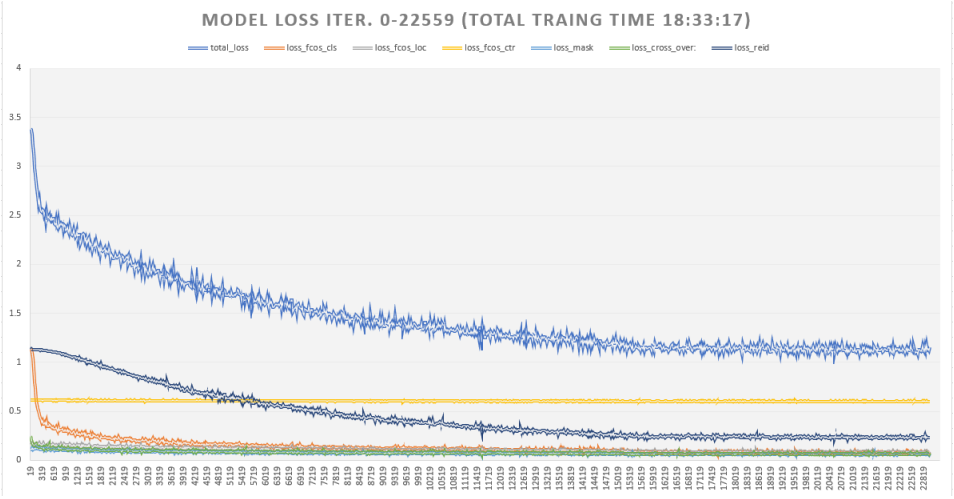

图 5：训练损失。。

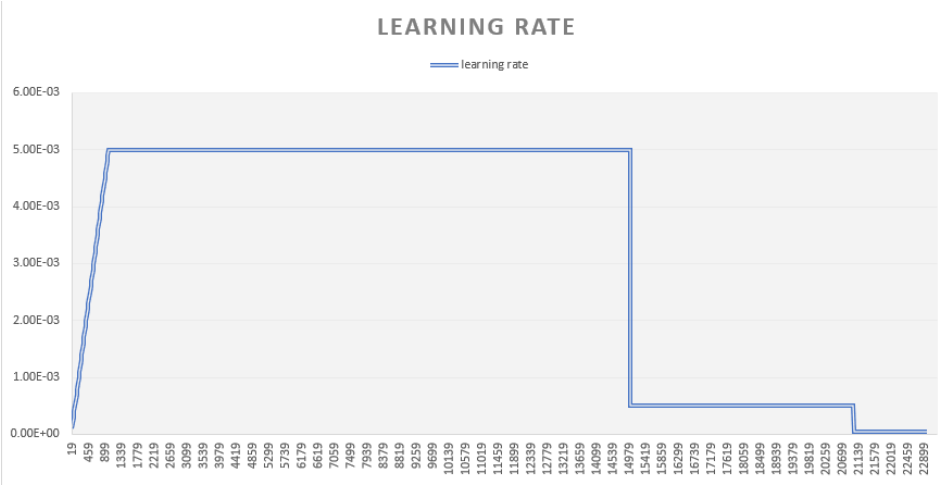

图 6：学习率。。

。我们想验证“Cross”结构是否确实能提升图像实例分割模型。因此，我们设计了一项移除“Cross”结构的实验。实验结果如图 7 所示。)

4.4.3 AudioFusion CrossVis性能

通过引入音频特征，我们构建了 Audio Fusion CrossVis 模型。训练后得到的损失如图 8 所示，可以看到，总损失和四项单独损失都如预期般下降。这表明模型已收敛，也证明了模型的正确性。随后，我们尝试比较 AFCV模型与传统CrossVis的收敛速度，性能如图 9 所示。起初，由于引入了音频这一新特征，损失高于基线。随着训练进行，我们模型的损失开始下降；经过约 12000 次迭代后，损失趋于稳定，并略低于基线。

在验证模型的正确性及其出色的收敛性能后，我们尝试使用评估指标，将其与 MaskTrack R-CNN、CondInst-VIS 和 CrossVis 等其他模型进行比较，并引入了两个数据集（COCO 2017 训练集和VIS数据集）。此外，我们采用平均精度（AP）作为评估指标来衡量性能。从表 1 可以看出，与 MaskTrack R-CNN 相比，由于模型深度较浅，我们的模型在实例分割方面表现较差。不过，在视频实例分割任务中，我们的模型

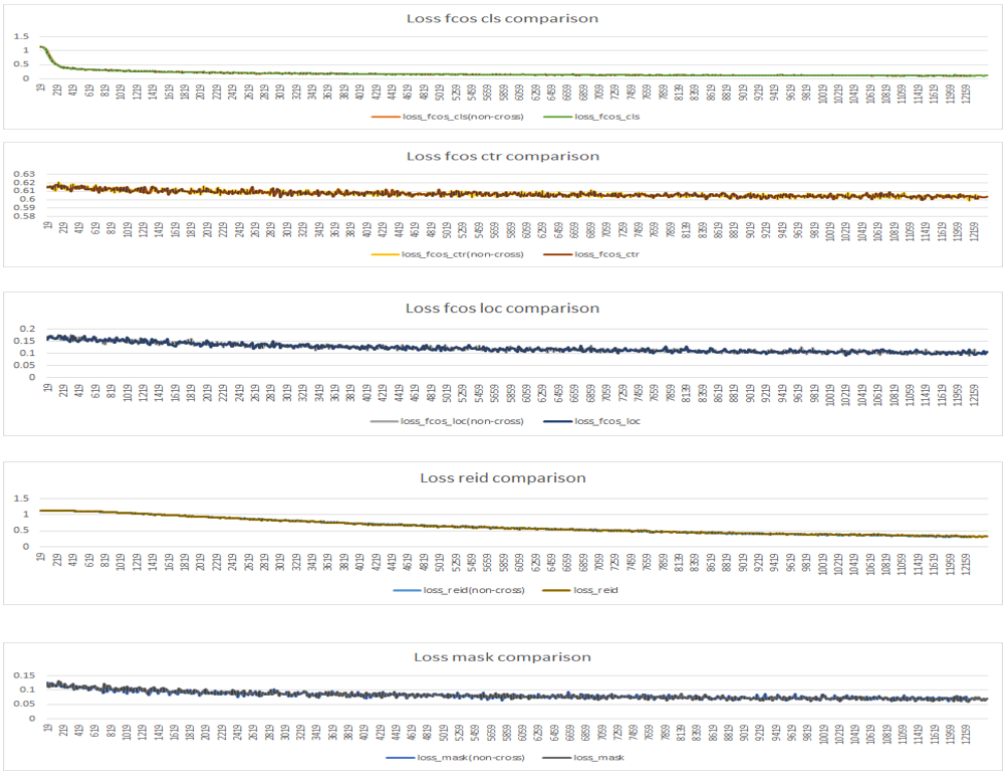

图 7：“CrossOver”移除实验。

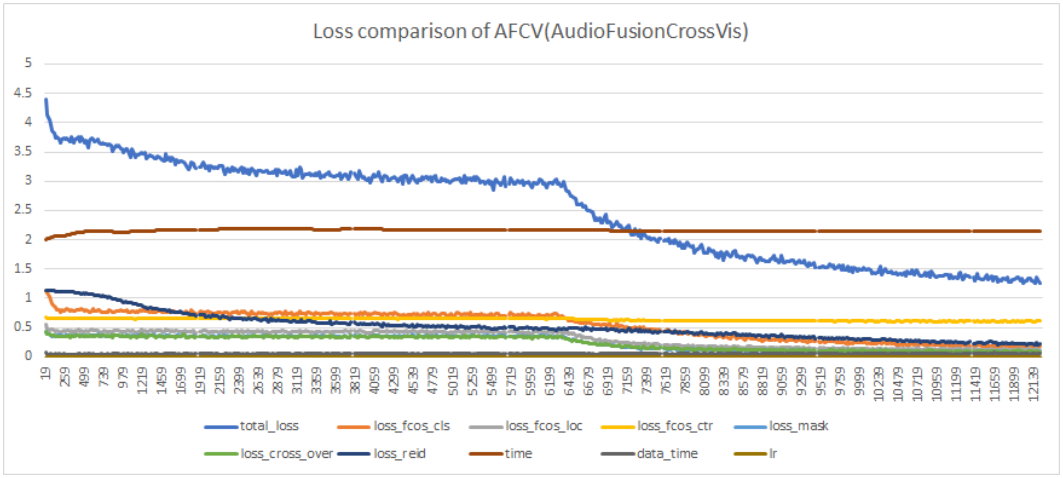

图 8：AFCV损失。

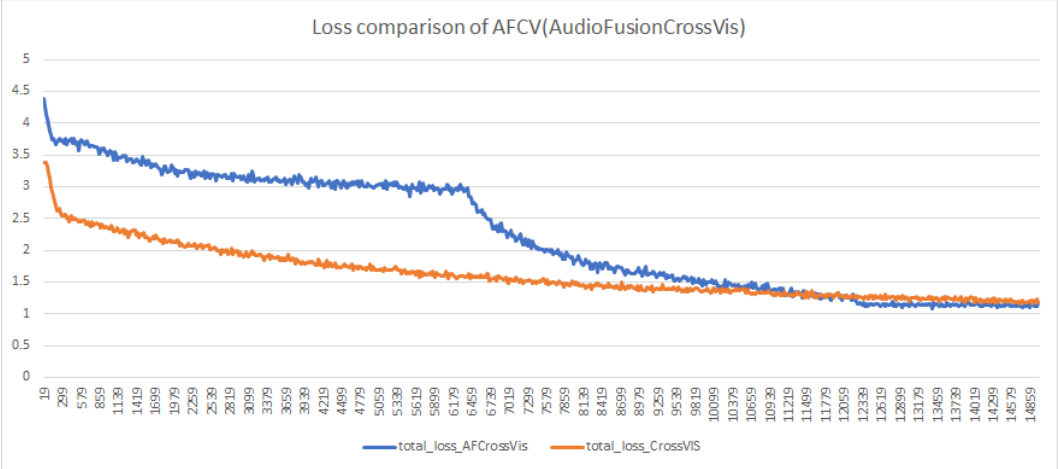

图 9：AFCV 收敛性能。

表 1： 评估指标

方法	骨干网络	平均精度 vis	平均精度 COCO MASK
MaskTrack RCNN[31]	ResNet-50	30.3	34.7
MaskTrack RCNN[31]	ResNet-101	31.9	35.9
CondInst-Vis	ResNet-50	32.1	35.7
CrossVis	ResNet-50	34.8	35.7
AFCV	ResNet-50	**34.9**	**35.7**

其表现令人鼓舞，平均精度高于基线模型和 MaskTrack R-CNN 模型。因此，我们可以得出结论，将音频特征融入模型是有效的。

5 结论

我们已成功复现结合在线视频实例分割方法的 CrossVis 框架解决方案，并进行了消融实验，以检验交叉学习对 CrossVIS 框架的有效性。然而，实验结果表明，与基线模型相比，交叉学习带来的性能提升有限。随后，鉴于视频来源中的音频通常未得到充分利用，我们将音频方面的输入融合到视觉数据中。通过使用组合解码器融合视频和音频特征，我们的模型相比基线模型表现出令人鼓舞的提升。未来，我们计划通过在 Sound-20K 数据集和自采集数据集中标注更多对象来扩大训练集规模。随后，我们将评估其他融合方法，并对缩放因子和宽高比进行敏感性分析。此外，我们还将实现在视听条件下的视觉注意力机制，以提高对声音生成区域的定位准确性。

参考文献

[1] S Avinash Ramakanth 和 R Venkatesh Babu。Seamseg：使用图像块接缝进行视频目标分割。发表于*IEEE*计算机视觉与模式识别会议论文集，第376–383页，2014年。[2] James M Bower 和 David Beeman。《*GENESIS* 之书》：使用通用神经模拟系统探索逼真的神经模型。施普林格科学与商业媒体，2012年。[3] Biagio Brattoli、Joseph Tighe、Fedor Zhdanov、Pietro Perona 和 Krzysztof Chalupka。重新思考零样本视频分类：面向实际应用的端到端训练。发表于论文集

*IEEE/CVF*计算机视觉与模式识别会议论文集，第4613–4623页，2020年。[4] Joshua Candamo、Matthew Shreve、Dmitry B Goldgof、Deborah B Sapper 和 Rangachar Kasturi。理解交通场景：人类行为识别算法综述。*IEEE*智能交通系统汇刊，11(1):206–224，2009年。[5] Jiale Cao、Rao Muhammad Anwer、Hisham Cholakkal、Fahad Shahbaz Khan、Yanwei Pang 和 Ling Shao。SipMask：为快速图像与视频实例分割保留空间信息。发表于计算机视觉—*ECCV 2020*：第*16*届欧洲会议，英国格拉斯哥，*2020*年*8*月*23*日至*28*日，论文集，第*XIV*部分*16*，第1–18页。施普林格，2020年。[6] William Chan、Navdeep Jaitly、Quoc Le 和 Oriol Vinyals。Listen, Attend and Spell：用于大词汇量会话语音识别的神经网络。发表于*2016*年*IEEE*国际声学、语音与信号处理会议（*ICASSP*），第4960–4964页。IEEE，2016年。[7] Honglie Chen、Weidi Xie、Andrea Vedaldi 和 Andrew Zisserman。VGGSound：大规模音视频数据集。发表于*ICASSP 2020—2020*年*IEEE*国际声学、语音与信号处理会议（*ICASSP*），第721–725页。IEEE，2020年。[8] 冯子舜、Ming Tu、Yuxuan Wang、Ashok Krishnamurthy 等。野外视频的自监督音视频表征学习。发表于*2020*年*IEEE*国际大数据会议（大数据），第5671–5672页。IEEE，2020年。[9] Yang Fu、Linjie Yang、Ding Liu、Thomas S Huang 和 Humphrey Shi。CompFeat：用于视频实例分割的全面特征聚合。*arXiv*预印本 *arXiv:2012.03400*，2020年。[10] Daniela Giordano、Francesca Murabito、Simone Palazzo 和 Concetto Spampinato。基于知觉组织和位置先验的超像素视频目标分割。发表于*IEEE*计算机视觉与模式识别会议论文集，第4814–4822页，2015年。[11] Matthias Grundmann、Vivek Kwatra、Mei Han 和 Irfan Essa。高效的基于层次图的视频分割。发表于*2010*年*IEEE*计算机学会计算机视觉与模式识别会议，第2141–2148页。IEEE，2010年。[12] Kaiming He、Georgia Gkioxari、Piotr Dollár 和 Ross Girshick。Mask R-CNN。发表于*IEEE*国际计算机视觉会议论文集，第2961–2969页，2017年。[13] Kaiming He、Xiangyu Zhang、Shaoqing Ren 和 Jian Sun。用于图像识别的深度残差学习。发表于*IEEE*计算机视觉与模式识别会议论文集，第770–778页，2016年。[14] Peiliang Huang、Junwei Han、Nian Liu、Jun Ren 和 Dingwen Zhang。涂鸦监督的视频目标分割。*IEEE/CAA*自动化学报，9(2):339–353，2021年。[15] Yong Huang、Xiang Li、Wei Wang、Tao Jiang 和 Qian Zhang。利用Wi-Fi信道状态信息检测监控视频流中的伪造攻击。*IEEE*无线通信汇刊，2021年。[16] Yong Huang、Xiang Li、Wei Wang、Tao Jiang 和 Qian Zhang。面向实时监控视频的跨模态伪造检测与定位。发表于*IEEE INFOCOM 2021—IEEE*计算机通信会议，第1–10页。IEEE，2021年。[17] Suyog Dutt Jain、Bo Xiong 和 Kristen Grauman。FusionSeg：学习结合运动与外观，对视频中的通用对象进行全自动分割。发表于*2017*年*IEEE*计算机视觉与模式识别会议（*CVPR*），第2117–2126页。IEEE，2017年。[18] Zihang Lai、Erika Lu 和 Weidi Xie。MaST：记忆增强型自监督跟踪器。发表于*IEEE/CVF*计算机视觉与模式识别会议论文集，第6479–6488页，2020年。[19] Minghan Li、Shuai Li、Lida Li 和 Lei Zhang。用于高效单阶段视频实例分割的空间特征校准与时间融合。发表于*IEEE/CVF*计算机视觉与模式识别会议论文集，第11215–11224页，2021年。[20] Xiang Li、Haoyuan Cao、Shijie Zhao、Junlin Li、Li Zhang 和 Bhiksha Raj。全景声引导的全景视频显著目标检测。*arXiv*预印本 *arXiv:2211.14419*，2022年。

[21] Xi'ang Li, Jinqi Luo, and Rabih Younes. ActivityGAN：用于基于传感器的人类活动识别数据增强的生成对抗网络。载于*2020*年*ACM*普适与泛在计算国际联合会议附属论文集及*2020*年*ACM*可穿戴计算国际研讨会论文集，第249–254页，2020年。[22] Xiang Li, Jinglu Wang, Xiao Li, and Yan Lu. 用于在线视频实例分割的混合实例感知时序融合。载于*AAAI*人工智能会议论文集，第36卷，第1429–1437页，2022年。[23] Xiang Li, Jinglu Wang, Xiao Li, and Yan Lu. 通过实例流组装实现视频实例分割。*IEEE*多媒体汇刊，2022年。[24] Xiang Li, Jinglu Wang, Xiaohao Xu, Bhiksha Raj, and Yan Lu. 通过鲁棒上下文融合实现在线视频实例分割。*arXiv*预印本 *arXiv:2207.05580*，2022年。[25] Jinqi Luo, Xiang Li, and Rabih Younes. 探索基于传感器的细粒度手部活动识别中的数据增强与解释。载于人类活动识别深度学习国际研讨会，第30–42页。施普林格，2021年。[26] Bruce McIntosh, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. 用于基于文本的视频分割的视觉-文本胶囊路由。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第9942–9951页，2020年。[27] Jie Pu, Yannis Panagakis, Stavros Petridis, Jie Shen, and Maja Pantic. 基于低秩与稀疏性的盲音视定位与分离。*IEEE*控制论汇刊，50(5):2288–2301，2018年。[28] Pavel Tokmakov, Karteek Alahari, and Cordelia Schmid. 利用视觉记忆学习视频目标分割。载于*IEEE*国际计算机视觉会议论文集，第4481–4490页，2017年。[29] Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. 将语言作为指代视频目标分割的查询。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第4974–4984页，2022年。[30] Linjie Yang, Yanran Wang, Xuehan Xiong, Jianchao Yang, and Aggelos K Katsaggelos. 通过网络调制实现高效的视频目标分割。载于*IEEE*计算机视觉与模式识别会议论文集，第6499–6507页，2018年。[31] Shusheng Yang, Yuxin Fang, Xinggang Wang, Yu Li, Chen Fang, Ying Shan, Bin Feng, and Wenyu Liu. 快速在线视频实例分割的交叉学习。载于*IEEE/CVF*国际计算机视觉会议论文集，第8043–8052页，2021年。[32] Neil Zeghidour, Olivier Teboul, Félix de Chaumont Quitry, and Marco Tagliasacchi. LEAF：用于音频分类的可学习前端。*arXiv*预印本 *arXiv:2101.08596*，2021年。[33] Dingwen Zhang, Junwei Han, Le Yang, and Dong Xu. SPFTN：用于在弱标注视频中定位和分割目标的联合学习框架。*IEEE*模式分析与机器智能汇刊，42(2):475–489，2018年。[34] Chenhui Zhao, Xiang Li, Siwen Dong, and Rabih Younes. 自监督多模态视频伪造攻击检测。*arXiv*预印本 *arXiv:2209.06345*，2022年。[35] Lingyu Zhu and Esa Rahtu. 利用类别信息进行单帧视觉声源分离。载于*2021*年第九届欧洲视觉信息处理研讨会（*EUVIP*），第1–6页。IEEE，2021年。[36] Dmitry Zotkin, Ramani Duraiswami, and Larry S Davis. 通过粒子滤波器实现多模态三维跟踪与事件检测。载于*IEEE*视频事件检测与识别研讨会论文集，第20–27页。IEEE，2001年。