Audio-Visual Segmentation

*Jinxing Zhou[1,2], *Jianyuan Wang[3], Jiayi Zhang[2,4], Weixuan Sun[2,3], Jing Zhang[3], Stan Birchfield[5], Dan Guo[1], Lingpeng Kong[6,7], Meng Wang[1], Yiran Zhong[2,7]

[1]Hefei University of Technology, [2]SenseTime Research, [3]Australian National University, [4]Beihang University, [5]NVIDIA, [6]The University of Hong Kong, [7]Shanghai Artificial Intelligence Laboratory

Introduction

Motivation

- Recent research on audio-visual learning mainly focuses on coarse-grained scene understanding, such as the audio-visual event classification and the patch-level sound source localization (SSL).
- We propose the audio-visual segmentation (AVS) task, which aims to explore the pixel-level audio-visual correspondence. The goal of AVS is to output a pixel-level segmentation map of the object(s) that produce sound at the time of the image frame (Figure 1).

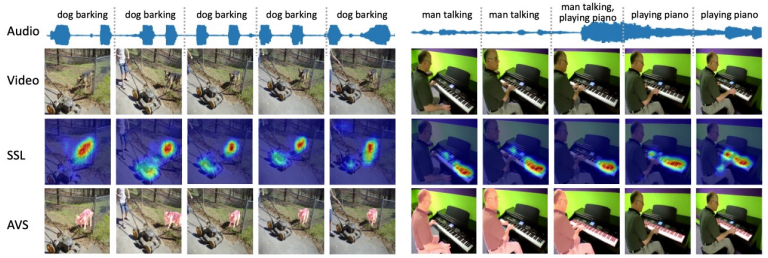


Figure 1. Comparison of the proposed AVS task with the SSL task.

AVSBench Dataset

To facilitate the research of AVS, we collect the AVSBench, a new dataset providing pixel-level annotations.

- **Data Statistics.** AVSBench contains 5,356 videos in total covering 23 types of common audio-visual scenes in real life. AVSBench can be divided into two subsets, i.e., the Single-source and the Multi-sources subset, according to the number of sound sources (Table 1).

subset	classes	videos	train/valid/test	annotated frames
Single-source	23	4,932	3,452*/740/740	10,852
Multi-sources	23	424	296/64/64	2,120

Table 1. AVSBench statistics.

- **Annotation.** For the Single-source subset, only the first sampled frame is annotated for the training process. As for the Multi-sources subset, all of the five sampled frames are annotated (Figure 2). This induces the two settings of the AVS task, i.e., the semi-supervised Single Sound Source Segmentation (S4) and the fully-supervised Multiple Sound Source Segmentation (MS3).

(a) Video examples in Single-source subset (b) Video examples in Multiple-sources subset

Figure 2. AVSBench samples.

Baseline Method

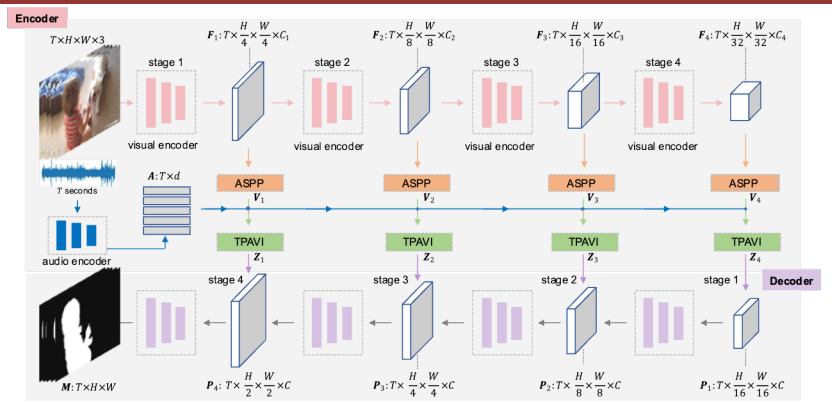


Figure 3. Overview of the Baseline Framework.

Baseline Encoder-Decoder Framework.

- The *encoder* takes the video frames and the entire audio clip as input, and outputs the visual and audio features, respectively denoted as $\boldsymbol{F}_i$ and $\boldsymbol{A}$. The visual feature map $\boldsymbol{F}_i$ at each stage is further sent to the ASPP [7] module and then our TPAVI module (Figure 4) which focuses on the temporal pixel-wise audio-visual interaction.
- The *decoder* progressively enlarges the fused feature maps by four stages and finally generates the output mask $\boldsymbol{M}$ for sounding objects.

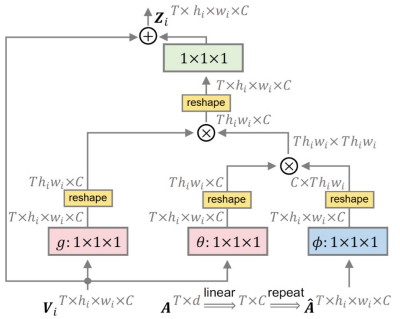


Figure 4. Illustration of the TPAVI module.

Objective Function. Given the prediction $\boldsymbol{M}$ and the pixel-wise label $\boldsymbol{Y}$, we adopt the binary cross entropy (BCE) loss as the main supervision functions.

Besides, we use an additional regularization term $\mathcal{L}_{\text{AVM}}$ to force the audio-visual mapping (Eq. 2). $\mathcal{L}_{\text{AVM}}$ computes the Kullback-Leibler (KL) divergence to ensure the masked visual features have similar distributions with the corresponding audio features.

$$\mathcal{L} = \text{BCE}(\boldsymbol{M}, \boldsymbol{Y}) + \lambda \mathcal{L}_{\text{AVM}}(\boldsymbol{M}, \boldsymbol{Z}, \boldsymbol{A}), \quad (1)$$

$$\mathcal{L}_{\text{AVM}} = \sum_{i=1}^{n} (\text{KL}(avg\ (\boldsymbol{M}_i \odot \boldsymbol{Z}_i), \boldsymbol{A}_i)), \quad (2)$$

Experiments

Comparison with methods from related tasks.

Metric	Setting	SSL		VOS		SOD		AVS (ours)	
		LVS[5]	MSSL[30]	3DC[27]	SST[10]	iGAN[28]	LGVT[49]	ResNet50	PVT-v2
$\mathcal{M}_{\mathcal{J}}$	S4	.379	.449	.571	.663	.616	.749	.728	**.787**
	MS3	.295	.261	.369	.426	.429	.407	.479	**.540**
$\mathcal{M}_{\mathcal{F}}$	S4	.510	.663	.759	.801	.778	.873	.848	**.879**
	MS3	.330	.363	.503	.572	.544	.593	.578	**.645**

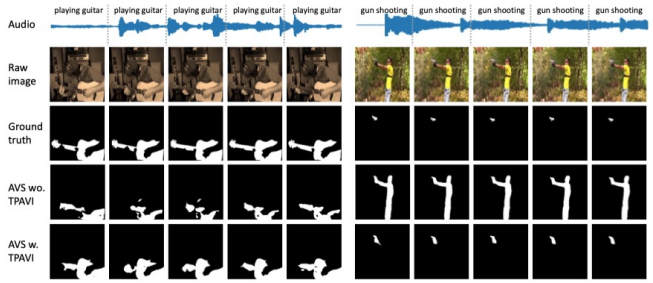
Impact of audio signal and the proposed TPAVI module.

AVS method	S4		MS3	
	ResNet50	PVT-v2	ResNet50	PVT-v2
without TPAVI	.701	.778	.436	.482
with A⊕V	.705	.777	.457	.516
with TPAVI	**.728**	**.787**	**.466**	**.531**

Effectiveness of the proposed $\mathcal{L}_{\text{AVM}}$.

Objective function	MS3 (mIoU)		MS3 (F-score)	
	ResNet50	PVT-v2	ResNet50	PVT-v2
$\mathcal{L}_{\text{BCE}}$	46.64	53.06	.558	.626
$\mathcal{L}_{\text{BCE}} + \mathcal{L}_{\text{AVM-VV}}$	46.71	53.77	.577	.644
$\mathcal{L}_{\text{BCE}} + \mathcal{L}_{\text{AVM-AV}}$	**47.88**	**54.00**	**.578**	**.645**

Qualitative results under the semi-supervised S4 setting.



Qualitative results under the fully-supervised MS3 setting.