Improving Audio-Visual Segmentation with Bidirectional Generation

Dawei Hao[1*], Yuxin Mao[2,3*], Bowen He[4], Xiaodong Han[2], Yuchao Dai[3], Yiran Zhong[2†]

[1]Bilibili Inc., Shanghai, China
[2]OpenNLPLab, Shanghai AI Lab, Shanghai, China
[3]Northwestern Polytechnical University, Shaanxi, China
[4]NIO, Shanghai, China
{howndawei, hannnnnkkf2511, daiyuchao, zhongyiran}@gmail.com,
maoyuxin@mail.nwpu.edu.cn, bowenhey@outlook.com

Abstract

The aim of audio-visual segmentation (AVS) is to precisely differentiate audible objects within videos down to the pixel level. Traditional approaches often tackle this challenge by combining information from various modalities, where the contribution of each modality is implicitly or explicitly modeled. Nevertheless, the interconnections between different modalities tend to be overlooked in audio-visual modeling. In this paper, inspired by the human ability to mentally simulate the sound of an object and its visual appearance, we introduce a bidirectional generation framework. This framework establishes robust correlations between an object's visual characteristics and its associated sound, thereby enhancing the performance of AVS. To achieve this, we employ a visual-to-audio projection component that reconstructs audio features from object segmentation masks and minimizes reconstruction errors. Moreover, recognizing that many sounds are linked to object movements, we introduce an implicit volumetric motion estimation module to handle temporal dynamics that may be challenging to capture using conventional optical flow methods. To showcase the effectiveness of our approach, we conduct comprehensive experiments and analyses on the widely recognized AVSBench benchmark. As a result, we establish a new state-of-the-art performance level in the AVS benchmark, particularly excelling in the challenging MS3 subset which involves segmenting multiple sound sources. Code is released in: https://github.com/OpenNLPLab/AVS-bidirectional.

Introduction

The foundation of human perception heavily relies on sight and hearing, which together absorb a substantial amount of external information. Integrating audio and visual information in a collaborative manner plays a crucial role in enhancing human scene understanding capabilities. Our daily experiences demonstrate that both auditory and visual cues contribute to our understanding of the concepts of objects. Therefore, audio-visual learning is essential in enabling machines to perceive the world through multi-modal information as humans do.

In the realm of audio-visual learning, the quest to disentangle auditory entities within videos at the minutiae of pixel level has given rise to the field of audio-visual segmentation (AVS). The pursuit of this goal has spurred the development of a myriad of techniques, often grounded in the fusion of multi-modal features, where the contribution of each modality is implicitly or explicitly modeled. However, a notable weakness in current methodologies is the insufficient attention given to the intricate relationship between various modes in audio-visual modeling. This lacuna forms the focal point of this paper.

Drawing inspiration from the remarkable human ability to conjure auditory perceptions corresponding to visual stimuli and vice versa, our study delves into a novel paradigm. However, directly replicating this procedure in a network, *i.e.,* generating audio from object segmentation masks, is difficult because the network needs to have audio creation capability, which requires a substantial amount of data to earn. Instead, we propose a bidirectional generation schema in feature space, meticulously designed to forge robust correlations between the visual attributes of objects and their corresponding auditory manifestations. Specifically, we utilize a visual-to-audio projection module to reconstruct audio features from object segmentation masks and minimize reconstruction errors. This schema allows the model to build a strong correlation between visual and audio signals. Furthermore, we recognize the profound relationship between sound and motion and introduce an implicit volumetric motion estimation module to address motions that may be difficult to capture using optical flow approaches (Zhong et al. 2019; Wang et al. 2020; Zhong et al. 2022). We construct a visual correlation pyramid for input video frames, make use of the inter-frame motion information to smooth the significant angle motion, and get clearer masks for dynamic objects.

To substantiate the efficacy of our novel framework, we undertake an extensive series of experiments and rigorous analyses on the AVSBench benchmark (Zhou et al. 2022). Our efforts culminate in the establishment of unprecedented benchmark performance, particularly within the challenging MS3 subset. In the spirit of scientific transparency and reproducibility, we commit to releasing both the code and the pre-trained models in the near future.

The key contributions of our work include:

*These authors contributed equally.
†Corresponding author.

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

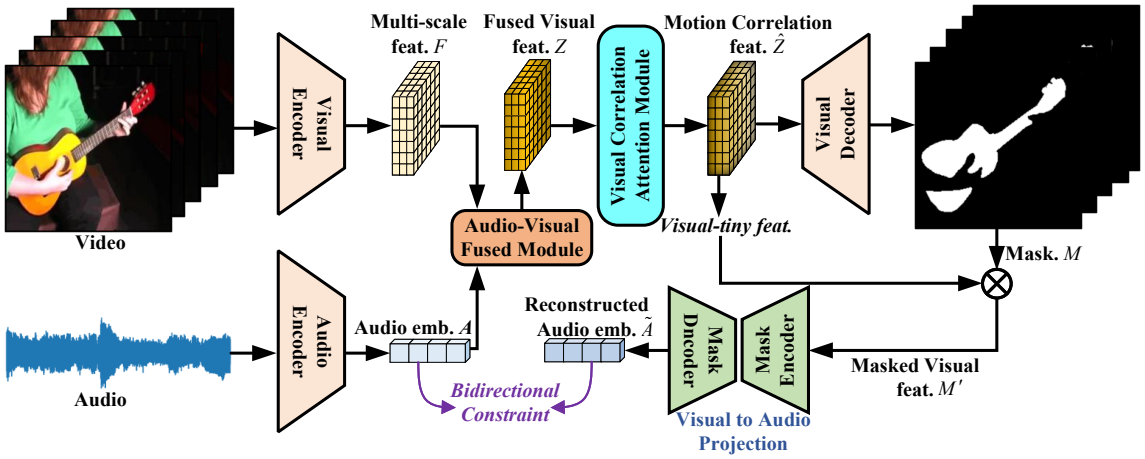


Figure 1: Overview of the proposed model, which follows a hierarchical Encoder-Decoder pipeline. The encoder inputs the video frames and the entire audio clip, and outputs multi-scale visual and audio features, respectively denoted as F and audio embedding A are further sent to the audio-visual fused module, which builds the audiovisual mapping to assist with identifying the sounding object. The visual correlation attention module predicts inter-frame movement information resulting in motion correlation features $\hat{Z}$. The decoder progressively enlarges the fused feature maps and finally generates the output mask M for sounding objects. In addition, masked visual features M' is supplied to visual-to-audio projection module, in which the result reconstructed audio embedding $\widetilde{A}$ and the original audio embedding A construct bidirectional constraint together.

- Proposed an innovative, efficient audio-visual segmentation approach using bidirectional generation supervision, which builds strong correlations between audio-visual modalities.
- Constructed a visual correlation pyramid for input video frames, leveraging implicit inter-frame motion information to enhance object mask quality.
- Achieved state-of-the-art performance on the AVSBench benchmark.

Related Work

Audio-Viusal Segmentation. More challenging than the sound source localization (SSL) (Chen et al. 2021; Cheng et al. 2020; Hu et al. 2020; Qian et al. 2020), event parsing (AVP) (Wu and Yang 2021; Tian, Li, and Xu 2020; Zhou et al. 2023) and event localization (AVEL) (Zhou et al. 2021; Zhou, Guo, and Wang 2022) task is audio-visual segmentation task. These methods require the fusion of audio and visual signals. Such as audio-visual similarity modeling by computing the correlation matrix (Arandjelovic and Zisserman 2017, 2018), audio-visual cross attention (Xuan et al. 2020), audio-guided Grad-CAM (Qian et al. 2020), or using a multi-modal transformer for modeling the long-range dependencies between elements across modalities directly. Zhou et al. (Zhou et al. 2022) released a segmentation dataset with pixel-level annotation of localization, and they introduced audio-visual segmentation (AVS). The audio-visual segmentation (Mao et al. 2023b,a) requires locating the actual sounding object(s) from multiple candidates and describing the contours of multiple sounding objects clearly and accurately.

Motion Estimation. To model motion information between video frames, a more traditional method is to explicitly use the method based on optical flow to model the pixel-by-pixel intensive correspondence between adjacent video frames (Tokmakov, Alahari, and Schmid 2017; Yang et al. 2021; Wang et al. 2020). However, for fast-moving objects and dynamic video scenes with occluded objects, the error of optical flow estimation accumulates, leading to the wrong analysis of motion information between two frames.

Bidirectional Consistency. In addition to modeling inter-frame motion estimates, our framework also incorporates cycle consistency as a supervisory signal. Hu, Chen, and Owens (2022) formulated the image and sound as graphs and adopted a cycle-consistent random walk strategy for separating and localizing mixed sounds. This method only locates the sounding area in videos but does not consider the shape of objects. In this paper, we introduce audio-visual-audio loop consistency constraints to build strong correlations between audio and segmented masks and strengthen audio supervision throughout the prediction process.

Method

In this section, we give a thorough explanation of the proposed method by first giving a general overview of the model's structure and then going into detail about each of its constituent parts.

The Overall Architecture

We illustrate the overall architecture of our proposed model in Figure 1. It consists of the following major parts: (1) Visual Encoder, (2) Audio Encoder, (3) Audio-Visual Fused Module, (4) Visual Correlation Attention Module, (5) Visual to Audio Projection Module, and (6) Visual Decoder.

Information Flow. The process begins with the encoder receiving video frames and the entire audio clip. It then outputs multi-scale visual and audio features, represented as F and A, respectively. These features are then sent to the audio-visual fused module, which creates an audio-visual map to aid in identifying the sounding object. The visual correlation attention module predicts inter-frame movement information to produce motion correlation features $\hat{Z}$. The decoder then progressively enlarges the fused feature maps to generate the output mask M for sounding objects. Additionally, masked visual features M' are provided to the visual-to-audio projection module, which uses them to reconstruct the audio embedding $\widetilde{A}$, while also considering the original audio embedding A to create a bidirectional constraint.

We define a video clip with a length of T as $\{\{I_t\}, \{A_t\}\}_{t=1}^{T}$, and the video frame $I_t \in \mathbb{R}^{3\times H\times W}$ corresponding to the audio A_t at the moment t as the reference frame, $\{A_t\}_{t=1}^{T}$ represents the audio with length T, and the pixel-level mask $\{M_t \in \{0,1\}\}_{t=1}^{T}$ corresponding to the network as the output. H, W represent the height and width of the frame, respectively.

Visual Encoder

We use Pyramid Vision Transformer (PVT-v2) (Wang et al. 2022) as our visual encoder. We load the pre-trained weights of ImageNet (Russakovsky et al. 2015) to speed up network convergence and improve performance. The input of our visual encoder is the T video frames $I \in \mathbb{R}^{T\times 3\times H\times W}$. The output is visual features of four scales $F_i \in \mathbb{R}^{T\times C_i\times h_i\times w_i}$, where $(h_i, w_i) = (H, W)\,/2^{i+1}$, $i = 1, 2, 3, 4$. C_i is feature channel and T represents the length of the video.

Audio Encoder

We first process $\{A_t\}_{t=1}^{T}$ to a Mel-Spectrogram via the short-time Fourier, and then fed it into VGGish (Hershey et al. 2017) that is pre-trained on AudioSet (Gemmeke et al. 2017) to obtain audio features $A \in \mathbb{R}^{T\times d}$, $d = 128$ is the feature dimension.

Audio-Visual Fused Module

We introduce a cross-modal fusion module for combining audio and video data, leveraging audio features to enhance vocal object prediction within video frames, as illustrated in Figure 2. Our approach involves several steps. Prior to fusion, we align the channel dimensions of multi-scale features extracted by the video encoder with those of both visual and audio features. This alignment yields transformed video features denoted as $F_i' \in \mathbb{R}^{T\times C\times h_i\times w_i}$, where $C = 128$ in our experimental setup. We then apply regularization and dimension transformation independently to the multi-scale video features $F_i{}'$ and audio features A, leading to $\hat{F}_i' \in \mathbb{R}^{T\times C\times h_i\times w_i}$ and $\hat{A}' \in \mathbb{R}^{T\times d\times 1\times 1}$, respectively.

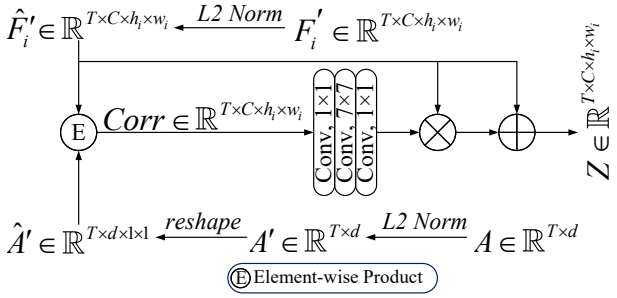


Figure 2: Audio-visual fused module takes the multi-scale features F_i' and audio features A as inputs. The symbols "$\oplus$" and "$\otimes$" denote matrix multiplication and element-wise addition, respectively.

Utilizing the Enisum method, we perform element-wise multiplication and matrix operations between $\hat{F_i}'$ and $\hat{A}'$ across the spatial dimensions (h, w), resulting in a correlation pyramid that captures the interplay between video and audio cues. This process culminates in the generation of fused video features denoted as $Z_i \in \mathbb{R}^{T\times C\times h_i\times w_i}$, as depicted below.

Visual Correlation Attention Module

To improve the accuracy of our predictions and reduce any disruptions caused by significant movements between adjacent frames, we must consider the motion information between these structures. Our solution is an implicit correlation pyramid that can model motion and predict segmentation simultaneously. This network can optimize both motions and segmentation with only segmentation supervision. However, using optical flow methods would require ground truth optical flow to supervise the optical flow module. Otherwise, it may lead to errors accumulating during the continuous prediction process, which could adversely affect segmentation performance.

By integrating the video features $Z_i \in \mathbb{R}^{T\times C\times h_i\times w_i}$ of four different scales after cross-modal fusion, we split the visual features ${Z_i}^t \in \mathbb{R}^{C\times h_i\times w_i}$ $(t = 1, \ldots, T)$ of the moment t according to the T dimension, and then construct the correlation pyramid $\widetilde{Z}_i^t \in \mathbb{R}^{h_i\times w_i\times h_i\times w_i}$ of two adjacent frames $\left\{{Z_i}^t, {Z_i}^{t+1}\right\}_{t=1}^{T}$. In the experiment, we constructed two cost calculations for the last two frames, and the calculation process is as follows:

$$\{\widetilde{Z}_i^t\}_{t=1}^{T} = \{\mathrm{Corr}({Z_i}^t, {Z_i}^{t+1}) \in \mathbb{R}^{h_i\times w_i\times h_i\times w_i}\}_{t=1}^{T} \quad (1)$$

By calculating the cost attention of visual features, the motion correlation features $\hat{Z}_i^t \in \mathbb{R}^{T\times C\times h_i\times w_i}$ are obtained.

Visual to Audio Projection

To build correlations between object appearance and its audio, we propose a visual to audio projection model to recon-

struct corresponding audio features according to the masked feature images.

We use the visual features $\hat{Z}_i$ obtained from the visual relevance attention module and select the minimum scale $\hat{Z}_4 \in \mathbb{R}^{T\times\frac{C}{4}\times h_4\times w_4}$ as a part of the reconstructed object, where $h_4 = w_4 = 7$. We obtain the masked feature map $M' \in \mathbb{R}^{T\times\frac{C}{4}\times 7\times 7}$ by calculating the correlation between the masked ground truth of actual sounding objects and the feature maps predicted by the segmentation network. We then input the reconstructed visual features M' into the mask Encoder and obtain $\hat{M'} \in \mathbb{R}^{T\times\frac{C}{4}\times 7\times 7}$. Finally, we input $\hat{M'}$ to the mask Decoder, which outputs the reconstructed audio features $\widetilde{A} \in \mathbb{R}^{T\times d\times 1\times 1}$. In our experiment, the Single Source subset serves as the predictive mask, while for the multi-source subset, M is the true mask.

Decoder

In order to predict the sound object mask in a video frame, we utilize the Decoder structure explained in (Zhou et al. 2022). The decoded features are subsequently upsampled to the following stage, and the final result of the decoder is $M \in \mathbb{R}^{T\times H\times W}$.

Loss Ojective

We adopt Binary cross entropy loss (BCE) and Kullback-Leibler (KL) divergence in (Zhou et al. 2022) as part of the objective function, as shown in the formula, where avg represents average pool operation and $\otimes$ is multiply operation element-by-element. The setting of hyper-parameter λ is similar to that of the paper (Zhou et al. 2022). During the training of S4, $\lambda = 0$. During the training in MS3, $\lambda = 0.5$.

$$\mathcal{L}_{\text{AVS}} = \text{BCE}\,(M, Y) + \lambda \sum_{i=0}^{n} \{\text{KL}[avg(M \otimes Z), A]\} \tag{2}$$

To ensure the consistency between the reconstructed audio features $\widetilde{A}$ and the original input audio A in the feature space, we designed latent loss $L_{\text{consistency}}$ to carry out consistency constraints on $\widetilde{A}$ and A, as shown in the Equ. 3.

$$\mathcal{L}_{\text{consistency}} = \eta\text{Distance}\left[\text{Norm}(\mathbf{A}), \text{Norm}(\widetilde{\mathbf{A}})\right] \tag{3}$$

In this paper, the expression calculates the norm along the last dimension, using the norm to calculate the regularized audio features $\widetilde{A}$ and A. We explored the influence of different hyperparameters η on the performance of the model. According to the experimental results in Table 2, we set the hyperparameter η equal to 1.0. In summary, the total loss functions can be written as follows:

$$\mathcal{L}_{\text{AVS-BG}} = \mathcal{L}_{\text{AVS}} + \mathcal{L}_{\text{consistency}} \tag{4}$$

Experiments Results

Implementation Details

Dataset. We conduct training and evaluation experiments on the AVSBench (Zhou et al. 2022) dataset. AVSBench contains two subsets, namely Single-source (S4) and Multi-sources (MS3), depending on the number of sounding objects(s). The single-source subset contains 4932 videos over 23 categories, covering sounds from humans, animals, vehicles, and musical instruments. The multi-source subset contains 424 videos and each video has multiple sounding sources and the sounding objects are visible in the frames. Each video was trimmed to 5 seconds. There are two settings of audio-visual segmentation: 1) semi-supervised Single Sound Source Segmentation (S4), and 2) fully supervised Multiple Sound Source Segmentation (MS3). For the Single-source set, only part of the ground truth is given during training (i.e., the first sampled frame of the videos) but all the video frames require a prediction during evaluation. In the Multi-sources subset, the labels of all five sampled

Tasks	Methods	S4		MS3	
		mIoU	F-score	mIoU	F-score
SSL	LVS	37.94	0.510	29.45	0.330
	MSSL	44.89	0.663	26.13	0.363
VOS	3DC	57.10	0.759	36.92	0.503
	SST	66.29	0.801	42.57	0.572
SOD	iGAN	61.59	0.778	42.89	0.544
	LGVT	74.94	0.873	40.71	0.593
AVS	AVS (R50)	72.79	0.848	47.88	0.578
	AVS (PVT)	78.74	0.879	54.00	0.645
	Ours (R50)	74.13	0.854	44.95	0.568
	Ours (PVT)	**81.71**	**0.904**	**55.10**	**0.668**

Table 1: Comparison with methods from related tasks. Results of mIoU and F-score under both S4 and MS3 are reported. We compare our method with LVS (Chen et al. 2021), MSSL (Qian et al. 2020), 3DC (Mahadevan et al. 2020), SST (Duke et al. 2021), iGAN (Mao et al. 2021), LGVT (Zhang et al. 2021), and AVS (Zhou et al. 2022).

Settings	S4		MS3 From Scratch	
	mIoU	F-score	mIoU	F-score
$\eta = 0.1$	81.40	0.900	53.07	0.648
$\eta = 0.5$	81.61	0.901	54.35	0.661
$\eta = 1.0$	**81.71**	**0.904**	**55.10**	**0.668**
$\eta = 1.5$	81.23	0.901	54.18	0.665

Table 2: Impact of the different hyperparameter η settings.

Variants	S4		MS3 From Scratch		MS3 Pretrained on Single-Source	
	mIoU	F-score	mIoU	F-score	mIoU	F-score
w/o audio	80.16	0.891	51.85	0.64	55.63	0.672
w audio	**80.82**	**0.895**	**54.35**	**0.661**	**56.51**	**0.682**

Table 3: Impact of the audio signal. Two rows back show the proposed method with or without the audio-visual fused module.

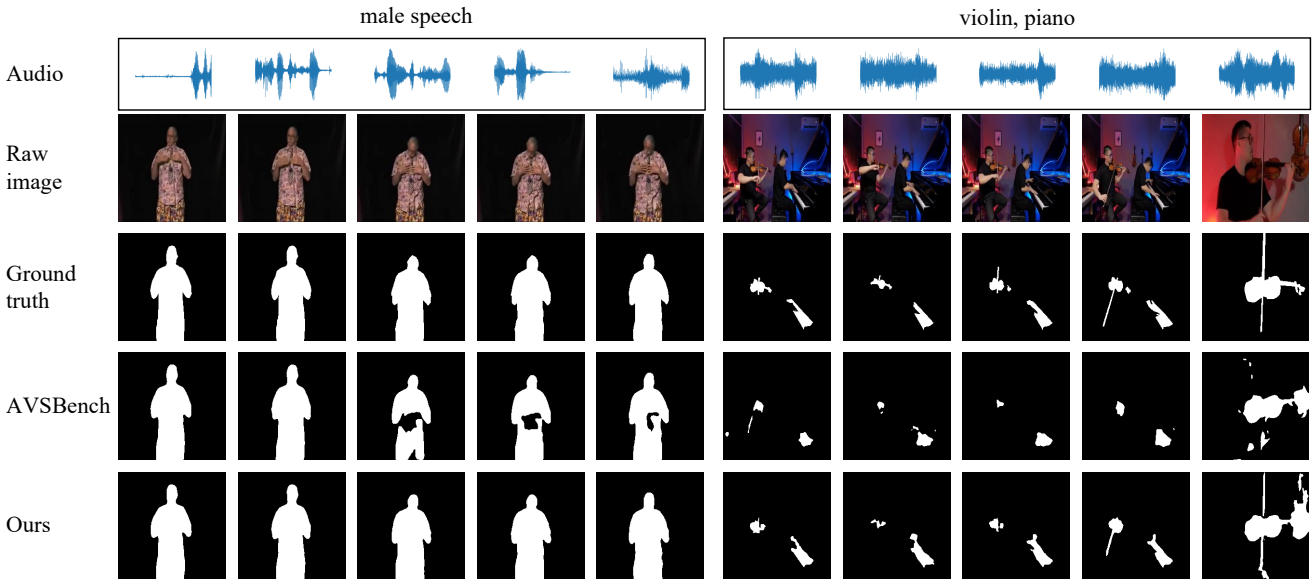


Figure 3: Qualitative examples of the AVSBench and our AVS framework. On the left is the output of the S4 setting, and on the right is the MS3 setting. AVSBench method only produces segmented maps that are not very precise, whereas our AVS framework can accurately segment the pixels of objects and create clear outlines of their shapes.

Variants	S4		MS3 From Scratch		MS3 Pretrained on Single-Source	
	mIoU	F-score	mIoU	F-score	mIoU	F-score
w/o motion	80.82	0.895	54.35	0.661	56.51	0.672
with motion	**81.26**	**0.898**	**54.91**	**0.670**	**58.33**	**0.697**

Table 4: Effectiveness of motion correlation.

frames of each are available for training. The goal for both settings is to correctly segment the sounding object(s) for each video clip by utilizing audio and visual cues.

Evaluation Metrics. We use the Mean Intersection-over-Union (mIoU) and F-score as the evaluation metrics. The former measures the contour accuracy of the predicted segmentation and the ground truth mask, and the latter considers both precision and recall. The F-score formulation is as follows, where β^2 is set to 0.3 in our experiments.

$$F_\beta = \frac{\left(1+\beta^2\right) \times \text{precision} \times \text{recall}}{\beta^2 \times \text{precision+recall}} \tag{5}$$

Quantitative Comparisons. Following the comparison settings of AVSBench (Zhou et al. 2022), we compare the performance of our method with the AVSBench baseline model and state-of-the-art methods from other related tasks, shown in Table 1. Our method consistently achieves significantly superior segmentation performance than the state-of-the-art methods, especially with 2.97 higher mIOU under the Single-source set than the previous AVSBench method. Our approach surpasses SSL, VOS, and SOD methods by a wide margin, indicating that the inclusion of audio data boosts segmentation accuracy.

Training Details. We train our model using PyTorch on an NVIDIA Tesla V100 and utilize the Adam optimizer with a learning rate of 10^{-4}. The batch size is set to 8, and we train on the Single-source subset for 15 epochs and the Multi-sources subset for 30 epochs. We resize all video frames to 224×224. Our experimental results are presented as percentages of the original results, and we use the PVT backbone to train all variations.

Comparison with Methods from Related Tasks

Qualitative Comparisons. We offer qualitative examples to compare our framework with previous AVSBench methods (Zhou et al. 2022). The quantitative results of the S4 setting and the MS3 setting are displayed in Figure 3. Our proposed model accurately segments all sounding objects and outlines their shapes with precision.

Ablation Studies

Impact of the Audio Signal. As illustrated in Figure 2, the audio-visual fused module is used for the audio-visual interactions from a temporal and pixel-wise level, introducing the audio information to explore the visual segmentation. We conduct an ablation study to explore its impact as shown in Table 3. We remove the audio part of the model and disable the audio-visual fused module to explore the importance of the audio information, leading to a simple unimodal framework with only video input, the results are shown in the first row of the table. For comparison, the second row shows the proposed method with the audio-visual fused module but

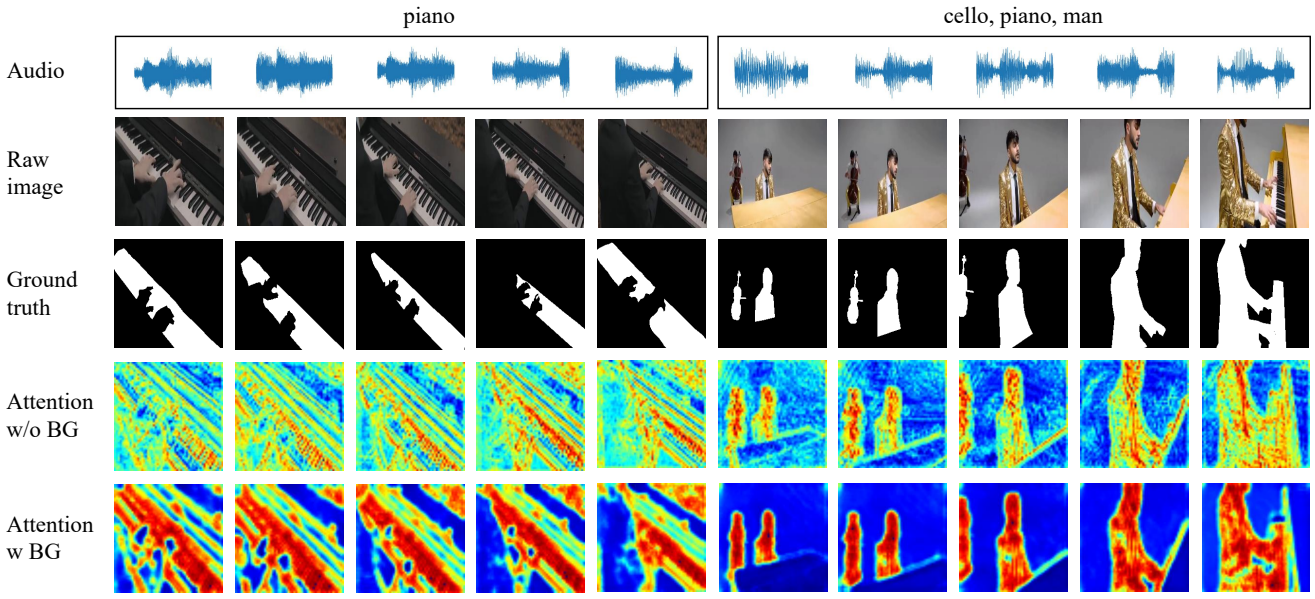


Figure 4: Audio-visual attention maps from motion correlation module under the multi-sources setting. A brighter color indicates a higher response. This indicates that our bidirectional generation constraint assists the model in focusing more on the visual regions that correspond to the audio.

without the motion correlation and bidirectional generation. It is noticed that adding the audio features to the visual one leads to a significant gain under the S4 and MS3 setting. This indicates directly the audio is especially beneficial to samples with single and multiple sound sources due to the audio signals can guide which object(s) to segment. Furthermore, our proposed audio-visual fused module can also help enhance the performance over various settings.

Effectiveness of Motion Correlation. We expect that the visual correlation module will estimate the movement information of intra-frames to enhance the network's accuracy to segment the correct objects. Therefore, we propose a motion correlation module. As shown in Table 4, smoothing big-angle movement achieves a clear performance gain. For example, Our method with motion correlation but without bidirectional generation improves the mIoU and F-score slightly. This demonstrates the benefits of introducing such a motion correlation.

Effectiveness of Bidirectional Generation. We conduct experiments to ablate the influence of cycle consistency constraints, as shown in Table 5. Under the dotted line, we reserve the motion estimation branch and explore the effectiveness of bidirectional generation. Although in the network framework with only a motion correlation module, our model already has a good segmentation performance. After introducing bidirectional generation, our model further improves the mIoU by around 0.25 and the F-score by about 0.018 in the MS3 setting pre-trained on the S4. Disabling motion estimation leads to even greater improvement when using bidirectional generation constraints.

To further verify the effectiveness of our proposed cycle-consistency constraint, we added bidirectional generation to the original AVSBench, and the experimental results are shown in Table 5. It is observed that the performance of the model can be significantly improved by using our proposed cycle consistency constraint while keeping the original AVSBench framework unchanged.

Besides, we also visualize the audio-visual attention matrices to explore what happens in the cycle-consistency constraint process. In detail, the attention map is obtained from the visual correlation attention module. We upsample it to have the same shape as the video frame. As shown in Figure 4, the high response area basically overlaps the region of sounding objects. It suggests that our bidirectional generation constraint builds a mapping from the visual pixels to the audio signals, which is semantically consistent.

Discussion of the Model Training and Inference

Comparison with Methods from Different Backbones. We compare the performance under different backbones with ResNet50 and PVT-v2. The performance on mIoU and F-score is reported in Table 1. It indicates that our framework has a significantly superior segmentation performance than the previous method. The performance gain comes from our designed visual correlation attention module and bidirectional generation constraint.

Pre-training on the Single-source Subset. Motivated by AVSBench (Zhou et al. 2022). The AVSBench model at the MS3 setting has been enhanced from 54.00 to 55.10. We conducted experiments by initializing model parameters

Methods		S4		MS3 From Scratch		MS3 Pretrained on Single-Source	
		mIoU	F-score	mIoU	F-score	mIoU	F-score
AVSBench	w/o BG	78.74	0.879	54.00	0.645	57.34	-
	with BG	**80.64**	**0.895**	**54.78**	**0.657**	**59.04**	**0.680**
Ours w/o motion	w/o BG	80.82	0.895	**54.35**	0.661	56.51	0.672
	with BG	**81.85**	**0.903**	53.49	**0.663**	**58.81**	**0.708**
Ours with motion	w/o BG	81.26	0.898	54.91	**0.670**	58.33	0.697
	with BG	**81.71**	**0.904**	**55.10**	0.668	**58.58**	**0.715**

Table 5: Effectiveness of bidirectional generation. We shorten bidirectional generation to BG. We conducted a comparison to see how much adding BG improved accuracy with and without our implicit motion estimation module. Our results show that BG greatly enhances accuracy when the motion estimation module is not present. However, when we combine motion and BG, our model performs even better. We also present the results of the AVSBench method with BG. These results demonstrate that our bidirectional generation approach is highly effective, regardless of the network structure.

Metrics	Setting	AVSBench		Ours	
		ResNet50	PVT-v2	ResNet50	PVT-v2
mIoU	S4	72.79	78.74	74.13	**81.71**
	MS3 From Scratch	47.88	54.00	44.95	**55.10**
	MS3 Pretrained on Single-Source	54.33	57.34	49.58	**58.58**
F-score	S4	0.848	0.879	0.854	**0.904**
	MS3 From Scratch	0.578	0.645	0.568	**0.668**
	MS3 Pretrained on Single-Source	-	-	0.627	**0.715**

Table 6: Comparison with methods from different backbones and different initialization strategies under the MS3 settings. We report the performance with ResNet50 and PVT-v2 as a backbone for the results of AVSBench and Ours under S4 and MS3 settings. We also report performance with different initialization strategies under the MS3 setting.

through pre-training on S4 dataset. The positive impact becomes more evident as the mIoU increases from 55.10 to 58.58 and the F-score increases from 0.668 to 0.715. Additionally, it has been proven that the pre-training strategy is advantageous in all settings.

Methods	Parameters (M)	Time (ms)
AVSBench	101.32	43.55
Ours	85.50	27.48

Table 7: Parameters and inference time. Our method achieves better accuracy while requiring fewer parameters and with a faster inference time.

Parameters and Inference Time. We have included a comparison of our parameters and inference time with AVSBench (Zhou et al. 2022) in Table 7. Our audio-visual fused module has replaced the TPAVI (Zhou et al. 2022) module, and we have reduced the number of neck channels from 256 to 128. This has resulted in a decrease in parameter numbers and a faster inference speed.

Conclusion

This paper presented a novel approach to audio-visual segmentation (AVS) that addresses the limitations of traditional methods by leveraging a bidirectional generation framework. By capitalizing on the human ability to mentally simulate the relationship between visual characteristics and associated sounds, our approach establishes robust correlations between these modalities, leading to enhanced AVS performance. The introduction of a visual-to-audio projection component, capable of reconstructing audio features from object segmentation masks, showcases the effectiveness of our methodology in capturing intricate audio-visual relationships. Additionally, the implicit volumetric motion estimation module tackles the challenge of temporal dynamics, particularly relevant for sound-object motion connections. Through comprehensive experiments and analyses conducted on the AVSBench benchmark, we demonstrated our approach's superiority, achieving a new state-of-the-art performance level, particularly excelling in complex scenarios involving multiple sound sources. Our work paves the way for further advancements in AVS. Future research could focus on refining the bidirectional generation framework, exploring novel ways to capture nuanced audio-visual associations, and investigating applications beyond AVS.

Acknowledgments

This research was supported in part by the National Natural Science Foundation of China (62271410), the National Key R&D Program of China (NO.2022ZD0160100), and the Fundamental Research Funds for the Central Universities.

References

Arandjelovic, R.; and Zisserman, A. 2017. Look, listen and learn. In *Proceedings of the IEEE international conference on computer vision*, 609–617.

Arandjelovic, R.; and Zisserman, A. 2018. Objects that sound. In *Proceedings of the European conference on computer vision (ECCV)*, 435–451.

Chen, H.; Xie, W.; Afouras, T.; Nagrani, A.; Vedaldi, A.; and Zisserman, A. 2021. Localizing visual sounds the hard way. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16867–16876.

Cheng, Y.; Wang, R.; Pan, Z.; Feng, R.; and Zhang, Y. 2020. Look, listen, and attend: Co-attention network for self-supervised audio-visual representation learning. In *Proceedings of the 28th ACM International Conference on Multimedia*, 3884–3892.

Duke, B.; Ahmed, A.; Wolf, C.; Aarabi, P.; and Taylor, G. W. 2021. Sstvos: Sparse spatiotemporal transformers for video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5912–5921.

Gemmeke, J. F.; Ellis, D. P.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R. C.; Plakal, M.; and Ritter, M. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 776–780. IEEE.

Hershey, S.; Chaudhuri, S.; Ellis, D. P.; Gemmeke, J. F.; Jansen, A.; Moore, R. C.; Plakal, M.; Platt, D.; Saurous, R. A.; Seybold, B.; et al. 2017. CNN architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, 131–135. IEEE.

Hu, D.; Qian, R.; Jiang, M.; Tan, X.; Wen, S.; Ding, E.; Lin, W.; and Dou, D. 2020. Discriminative sounding objects localization via self-supervised audiovisual matching. *Advances in Neural Information Processing Systems*, 33: 10077–10087.

Hu, X.; Chen, Z.; and Owens, A. 2022. Mix and localize: Localizing sound sources in mixtures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10483–10492.

Mahadevan, S.; Athar, A.; Ošep, A.; Hennen, S.; Leal-Taixé, L.; and Leibe, B. 2020. Making a case for 3d convolutions for object segmentation in videos. *arXiv preprint arXiv:2008.11516*.

Mao, Y.; Zhang, J.; Wan, Z.; Dai, Y.; Li, A.; Lv, Y.; Tian, X.; Fan, D.-P.; and Barnes, N. 2021. Transformer transforms salient object detection and camouflaged object detection. *arXiv preprint arXiv:2104.10127*.

Mao, Y.; Zhang, J.; Xiang, M.; Lv, Y.; Zhong, Y.; and Dai, Y. 2023a. Contrastive conditional latent diffusion for audio-visual segmentation. *arXiv preprint arXiv:2307.16579*.

Mao, Y.; Zhang, J.; Xiang, M.; Zhong, Y.; and Dai, Y. 2023b. Multimodal variational auto-encoder based audio-visual segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 954–965.

Qian, R.; Hu, D.; Dinkel, H.; Wu, M.; Xu, N.; and Lin, W. 2020. Multiple sound sources localization from coarse to fine. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, 292–308. Springer.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115: 211–252.

Tian, Y.; Li, D.; and Xu, C. 2020. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, 436–454. Springer.

Tokmakov, P.; Alahari, K.; and Schmid, C. 2017. Learning motion patterns in videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3386–3394.

Wang, J.; Zhong, Y.; Dai, Y.; Zhang, K.; Ji, P.; and Li, H. 2020. Displacement-invariant matching cost learning for accurate optical flow estimation. *Advances in Neural Information Processing Systems*, 33: 15220–15231.

Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; and Shao, L. 2022. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3): 415–424.

Wu, Y.; and Yang, Y. 2021. Exploring heterogeneous clues for weakly-supervised audio-visual video parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1326–1335.

Xuan, H.; Zhang, Z.; Chen, S.; Yang, J.; and Yan, Y. 2020. Cross-modal attention network for temporal inconsistent audio-visual event localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 279–286.

Yang, C.; Lamdouar, H.; Lu, E.; Zisserman, A.; and Xie, W. 2021. Self-supervised video object segmentation by motion grouping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7177–7188.

Zhang, J.; Xie, J.; Barnes, N.; and Li, P. 2021. Learning generative vision transformer with energy-based latent space for saliency prediction. *Advances in Neural Information Processing Systems*, 34: 15448–15463.

Zhong, Y.; Ji, P.; Wang, J.; Dai, Y.; and Li, H. 2019. Unsupervised deep epipolar flow for stationary or dynamic scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12095–12104.

Zhong, Y.; Loop, C.; Byeon, W.; Birchfield, S.; Dai, Y.; Zhang, K.; Kamenev, A.; Breuel, T.; Li, H.; and Kautz, J. 2022. Displacement-Invariant Cost Computation for Stereo Matching. *International Journal of Computer Vision*, 130(5): 1196–1209.

Zhou, J.; Guo, D.; and Wang, M. 2022. Contrastive positive sample propagation along the audio-visual event line. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhou, J.; Guo, D.; Zhong, Y.; and Wang, M. 2023. Improving audio-visual video parsing with pseudo visual labels. *arXiv preprint arXiv:2303.02344*.

Zhou, J.; Wang, J.; Zhang, J.; Sun, W.; Zhang, J.; Birchfield, S.; Guo, D.; Kong, L.; Wang, M.; and Zhong, Y. 2022. Audio–Visual Segmentation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII*, 386–403. f.

Zhou, J.; Zheng, L.; Zhong, Y.; Hao, S.; and Wang, M. 2021. Positive sample propagation along the audio-visual event line. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8436–8444.