석 사 학 위 논 문
Master's Thesis

SAM 기반 비디오 음원 위치 추정을 위한 시공간 상시청각 양방향 어텐션

SAM-based Audio-Visual Segmentation with Spatio-Temporal, Bidirectional Audio-Visual Attention

2024

선 주 형 (宣 主 形 Seon, Juhyeong)

한 국 과 학 기 술 원

Korea Advanced Institute of Science and Technology

석 사 학 위 논 문

SAM 기반 비디오 음원 위치 추정을 위한 시공간 상 시청각 양방향 어텐션

2024

선 주 형

한 국 과 학 기 술 원

전산학부

SAM 기반 비디오 음원 위치 추정을 위한 시공간 상 시청각 양방향 어텐션

선 주 형

위 논문은 한국과학기술원 석사학위논문으로
학위논문 심사위원회의 심사를 통과하였음

2024년 06월 05일

심사위원장 윤 성 의 (인)

심 사 위 원 홍 승 훈 (인)

심 사 위 원 안 성 진 (인)

SAM-based Audio-Visual Segmentation with Spatio-Temporal, Bidirectional Audio-Visual Attention

Juhyeong Seon

Advisor: Sung-Eui Yoon

A dissertation submitted to the faculty of
Korea Advanced Institute of Science and Technology in
partial fulfillment of the requirements for the degree of
Master of Science in Computer Science

Daejeon, Korea
June 05, 2024

Approved by

Sung-Eui Yoon
Professor of Computer Science

The study was conducted in accordance with Code of Research Ethics[1].

[1] Declaration of Ethical Conduct in Research: I, as a graduate student of Korea Advanced Institute of Science and Technology, hereby declare that I have not committed any act that may damage the credibility of my research. This includes, but is not limited to, falsification, thesis written by someone else, distortion of research findings, and plagiarism. I confirm that my thesis contains honest conclusions based on my own careful research under the guidance of my advisor.

MCS 선주형. SAM 기반 비디오 음원 위치 추정을 위한 시공간 상 시청각 양방향 어텐션. 전산학부 . 2024년. 22+iv 쪽. 지도교수: 윤성의. (영문 논문)
Juhyeong Seon. SAM-based Audio-Visual Segmentation with Spatio-Temporal, Bidirectional Audio-Visual Attention. School of Computing . 2024. 22+iv pages. Advisor: Sung-Eui Yoon. (Text in English)

초 록
비디오 음원 위치 추정 문제는 프레임 상에서 소리의 음원에 해당하는 물체의 위치를 예측하는 문제로, 이는 픽셀 수준에서 오디오와 시각 요소 간의 연결에 대한 자세한 이해를 요구합니다. 최근 발표된 Segment Anything Model (SAM)이 다양한 이미지 분할 문제에서 우수한 성능을 보임에 따라, SAM에 오디오 프롬프트를 도입하여 음원 위치 추정 문제를 해결하는 방법이 제안되습니다. 그러나 이러한 방법들은 모두 SAM의 단일 이미지 분할 기능에 기반하여 작동하기 때문에, 비디오 데이터에 표현된 여러 프레임 간의 관계 정보를 충분히 활용하지 못합니다. 이에 대응하여, 본 연구는 SAM이 복수 프레임 간의 관계를 활용하여 소리가 있는 비디오 장면에 대한 맥락적 이해를 높이고 음원 위치 추정의 높은 성능을 달성하는 방안을 탐구합니다. 이를 달성하기 위해 시공간적 양방향 교차 모달리티 어텐션 모듈을 도입하여 오디오-이미지 특성을 정제하고 비디오 프레임과 해당 오디오 트랙 간의 시공간 정합성을 이해하도록 합니다. 실험을 통해 제안된 모델이 음원 위치 추정 벤치마크에서 최신 방법들을 능가함을 보였으며, 특히 다중 음원이 등장하는 벤치마크에서 8.3%의 mIoU 향상을 달성함을 보입니다.

핵 심 낱 말 시청각 학습, 음원 위치 추정, 시청각 영상 음원 위치 분할, Segment Anything, 시공간 어텐션

Abstract
Audio-visual segmentation (AVS) focuses on segmenting sound sources within video sequences, necessitating a detailed understanding of the connections between audio and visual elements at the pixel level. As the Segment Anything Model (SAM) has profoundly influenced various dense prediction challenges, recent research has focused on integrating SAM into AVS with audio prompts. However, SAM's design primarily addresses single-frame segmentation, which does not fully leverage the temporal dynamics inherent in audio-visual data. To address this, our study explores enhancing SAM's functionality to encompass sequences of audio-visual scenes by examining the intermodal relationships across frames. We introduce a Spatio-Temporal, Bidirectional Audio-Visual Attention (ST-BAVA) module to dynamically refine audio-visual features and reinforce the spatio-temporal alignment between video frames and corresponding audio tracks. Extensive experiments demonstrate that our proposed model outperforms the state-of-the-art methods on AVS benchmarks, especially with an 8.3% mIoU gain on a challenging multi-sources subset.

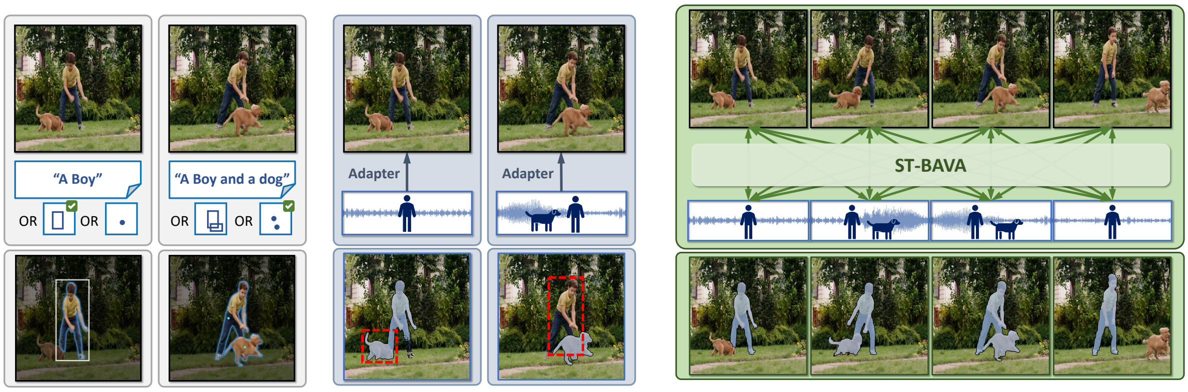
Keywords Audio-visual learning, sound source localization, audio-visual segmentation, Segment Anything, spatio-temporal attention

Contents

List of Tables

List of Figures

Chapter 1. Introduction



(a) SAM (b) Prior approaches [2, 3] (c) SAM with ST-BAVA (Ours)

Figure 1.1: Segmentation results of different models on a film *A Dog's Purpose* (2017). (a) Segment Anything Model (SAM) segments target objects in the image with their regions guided by user prompts. (b) Prior works [2, 3] have adapted SAM to segment objects that sound with a corresponding audio prompt per frame. (c) We propose a spatio-temporal, bidirectional audio-visual attention (ST-BAVA), enabling SAM to fully leverage the relationships between the subsequent video frames and audio streams in a bidirectional way. In Fig. 1.1c, our model successfully segments the human and the dog on the frames where they make sounds.

Segment Anything Model (SAM) [4] is a foundation model in image segmentation with points, boxes, and text prompts (Fig. 1.1a). Tremendous works have shown SAM's outstanding performance in various dense prediction problems [5, 6, 7, 8, 9, 10, 11, 12, 13] by adapting it to specific domains. For instance, Cen *et al.* [5] extended the segmentation ability of SAM to 3D scenes through NeRFs. Wu *et al.* [7] adapted SAM into the medical domain, showing generalized segmentation performance on CT, MRI, ultrasound, fundus, and dermoscopic images.

In light of the success of SAM in various tasks, pioneering works [14, 2, 3] have attempted to introduce a prompt in the novel modality, audio, into the SAM (Fig. 1.1b). They aim to segment the object that makes the sound in the audible video, defined as an Audio-Visual Segmentation (AVS) problem [1]. Mo *et al.* [14] explored a spatial fusion of the audio-visual features for audio prompting of SAM. Liu *et al.* [2] and Wang *et al.* [3] presented prompt tuning techniques [15] by inserting the light-weight adapters into the image encoder and decoder of SAM respectively, achieving high performance on the AVS benchmark.

Nevertheless, since they divide the video into individual frames in a single-frame manner constrained by SAM, the contextual information provided by the audio-visual scene across subsequent frames has been neglected. Furthermore, while the effectiveness of bidirectional modeling with the interaction between the image frame and the corresponding prompt has been demonstrated in the SAM-based referring video object segmentation (RVOS) [12] and tracking [13], it has not been sufficiently explored for the audible video data. As the AVS requires comprehensive pixel-level correspondence across two different modalities, leveraging the complementary relationship between audio-visual cues in a bidirectional way becomes more essential to this task.

To this end, we study the extension of SAM's segmentation capabilities to the subsequent frames with corresponding audio, proposing a Spatio-Temporal, Bidirectional Audio-Visual Attention module (ST-BAVA) (Fig. 1.1c). It aims to convey the spatio-temporal relationships between input images and audio prompts to SAM through bidirectional adjustments of audio-visual features in the middle of the pipeline. Our proposed method exhibits two advantages compared to the previous approaches [2, 14, 3]. First, we enable SAM to leverage the contextual information presented in the audio-visual scene of multiple frames. Second, we mutually align the image and audio features based on their spatial and temporal correspondences across the video. Through the bidirectional aggregation of the spatio-temporal relationship across the sequence of the audio-visual frames, our model effectively identifies the distinct visual and auditory cues from the objects that could emerge or disappear. Extensive experimental results (Sec. 5, 6) demonstrate that the proposed method achieves high performance in localizing the sound source with only a few trainable parameters (<4% of SAM), thanks to the audio-visual relationship exploited by ST-BAVA. In particular, it achieves improvements of 8.3% in mIoU on the AVS benchmark's multi-sources subset, as shown in Table 6.1.

We summarize our contributions as follows:

- We extend SAM into the auditory and temporal dimensions to segment the sound sources on the subsequent video frames using corresponding audio as a prompt.
- We propose a Spatio-Temporal, Bidirectional Audio-Visual Attention module (ST-BAVA), enabling SAM to exploit the spatio-temporal correspondences between subsequent image frames and audio streams.
- Through experiments, we demonstrate that the proposed method outperforms the state-of-the-art methods on the AVS. Furthermore, we showcase the effectiveness of the main components in our approach through extensive ablation studies.

Chapter 2. Related Work

2.1 Audio-Visual Segmentation

In audio-visual learning, the relationship between audio and visual data has been explored to understand the scene in multimodal. Researchers pioneered the audio-visual correspondence (AVC) [16, 17] with a binary classification to predict whether the image and audio data correspond. The relationship exploited in AVC has been extended into the spatial dimension, evolving into the task named sound source localization (SSL) that aims to localize the region of sound sources in the frame [18, 19, 20]. With a remarkable achievement of SSL, Zhou *et al.* [1] introduced an advanced segmentation challenge called Audio-Visual Segmentation (AVS). Diverse approaches have focused on learning effective multimodal representations to comprehend the pixel-level audio-visual correspondences in the video scene. Mao *et al.* [21, 22] introduced the multimodal VAE and conditional latent diffusion model to learn the advanced audio-visual representation. Huang *et al.* [23] and Gao *et al.* [24] utilized the transformer-based architecture with the interaction between audio-conditioned object queries and visual features. Liu *et al.* [25] proposed the two-stage framework: segmenting all potential objects from the visual data and verifying sounding objects using an audio-visual semantic matching. In this work, we introduce a bidirectional cross-modal feature interaction module to extend the capabilities of the Segment Anything Model (SAM) [4] to the AVS.

2.2 Segment Anything Model

SAM is a foundation model for generality and broad applications in image segmentation problems with point, box, mask, and text prompts, pre-trained on 1B masks from 11M images [4]. Extensive works have studied the SAM's ability with various problems, including 3D vision tasks [5, 6], medical image segmentation [7, 8], and shadow detection [9, 10]. Within this context, the extension of SAM to handle the video data has been explored [13, 12, 11]. For instance, Cheng *et al.* [11] proposed a user-interactive video object tracking framework by supporting SAM's insufficient temporal and semantic understanding of the object. They employed Grounding Dino [26], a vision-language model, to interactively convert the user's description into the box prompt for SAM. Beyond these approaches, our work introduces SAM's temporal extension with audio as a new input modality without relying on the existing prompts.

Chapter 3. Preliminary

In this chapter, we formulate the audio-visual segmentation problem, where the model predicts the pixel-level segmentation mask of sounding objects on the video frames with the corresponding audio prompts (Sec. 3.1). Also, we revisit the image segmentation process of SAM as a baseline of our method (Sec 3.2).

3.1 Problem Formulation

The video data for the audio-visual segmentation consists of a series of visual frames and audio spectrograms. For the video n with a length of T seconds, denoted as (I^n, S^n), the image frames are represented as $I^n = \{i_t^n\}_{t=1}^T$ with an image $i_t^n \in \mathbb{R}^{3\times H_i \times W_i}$ at timestep t. Each video frame is extracted at the end of each second. The audio spectrograms are represented as $S^n = \{s_t^n\}_{t=1}^T$ with $s_t^n \in \mathbb{R}^{H_s \times W_s}$. The spectrogram is processed via a short-time Fourier transform of the 1-second audio clip. The model outputs the segmentation map of the sound source as a binary mask $Y^n = \{y_t^n\}_{t=1}^T$ with $y_t^n \in \{0,1\}^{H_i \times W_i}$. Each pixel in y_t^n represents whether it is a sounding object. Multiple sound sources can be depicted within the single mask y_t^n without considering semantic differences for each object. The target segmentation mask is also provided as a binary mask $M^n = \{m_t^n\}_{t=1}^T$ with $m_t^n \in \{0,1\}^{H_i \times W_i}$. For simplicity, we omit the notation n afterward.

3.2 Revisiting SAM

SAM is an architecture designed to solve the image segmentation problem. It has three main components: an image encoder, a prompt encoder, and a mask decoder. For the image encoder, a Vision Transformer (ViT) [27] pre-trained with MAE [28] is used to extract the spatial image features. A prompt encoder supports two subsets of prompts: sparse (points, boxes, and text) and dense prompts (masks). Sparse prompts are embedded by the shallow encoder, and dense embedding is summed with the image embedding. These embeddings are fed into the mask decoder, providing the regional information about the target objects to segment.

We elaborate on the operation of the SAM decoder, which is essential to guiding the target's location through user prompts (Fig. 3.1). The decoder block uses prompt self-attention and bidirectional cross-attention to update both prompt and image embedding. The decoder's cross-attentions capture the spatial correspondence of the prompt and visual features at each time step to update both. The final similarity map between the fused image embedding and the prompt token is used as a final mask prediction of the target object.

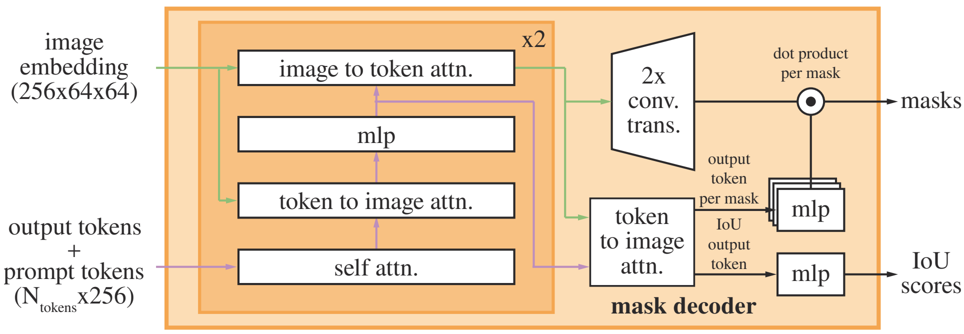


Figure 3.1: Details of the SAM's lightweight mask decoder from the paper by Kirillov *et al.* [4].

Chapter 4. Method

4.1 Extending SAM into Auditory and Temporal Dimensions

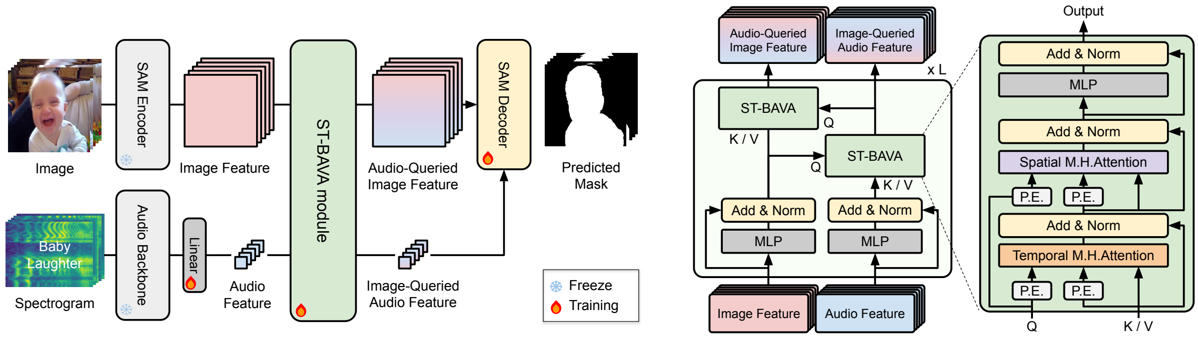
While SAM utilizes visual prompts to address various image segmentation problems, enabling SAM to support the audio prompts within multi-frame video requires an alternative approach. In the pipeline of SAM, the guidance of prompts about where to segment is done in the lightweight mask decoder with cross attention between the image and prompt embeddings as elaborated in Sec. 3.2. Considering this, a straightforward approach to processing the audio in SAM is to forward the audio embeddings as the prompt into the decoder, as introduced by Liu *et al.* [2]. However, using audio to represent the target objects in images is more complicated than using points and boxes, making it challenging for the lightweight decoder to comprehend the intricate audio-visual relationship [2]. Moreover, as the decoder operates the cross attention per frame, the temporal dependencies among subsequent frames are insufficiently leveraged. To solve these challenges, we propose the intermediate audio-visual interaction module ST-BAVA before the decoder. This module aggregates the spatio-temporal relationship of the audio-visual features extracted from the subsequent video frames and audio streams. These updated features by ST-BAVA are handed to the subsequent decoding process of the segmentation map, enabling the SAM's decoder to utilize the audio-visual correspondence in spatial and temporal dimensions.

4.2 SAM with ST-BAVA

Fig. 4.1a shows the pipeline of SAM with the proposed ST-BAVA module. SAM's image encoder embeds input images I to get the visual embedding $V \in \mathbb{R}^{T \times HW \times C}$, where H, W represent the spatial size of the visual embeddings and C represents a channel dimension. The audio backbone encoder, followed by the learnable linear layer, encodes the audio spectrograms S to the audio embedding $A \in \mathbb{R}^{T \times C}$, aligning with the channel dimension of visual embedding. V and A are forwarded to the ST-BAVA, where the spatial and temporal attention operates sequentially. Spatial attention computes the spatial correspondence between V and A for each time step, and temporal attention analyzes their audio-visual correlation across consecutive frames per pixel. By the bidirectional operation of the spatial and temporal attention, ST-BAVA produces the audio-queried visual embedding V_{aq} and visual-queried audio embedding A_{vq}. Both are inserted into the SAM's mask decoder where V_{aq} serves as the dense prompt and A_{vq} as the sparse prompt, replacing the original point or box prompts. The following decoding steps are carried out in the same way as in image segmentation. As the objective function for the AVS, we adopt the binary cross-entropy loss with the ground truth pixel mask M for prediction Y, represented as $\mathcal{L} = \text{BCE}(Y, M)$.

4.3 Spatio-Temporal, Bidirectional Audio-Visual Attention

Fig. 4.1b elaborates ST-BAVA module architecture. The design of ST-BAVA is motivated by the SAM decoder's bidirectional image-prompt feature fusion, extending it into spatio-temporal dimensions between the sequence of image and audio embeddings. As processing audible video data requires high



(a) Pipeline of SAM with ST-BAVA

(b) ST-BAVA module architecture

Figure 4.1: Overview of the proposed SAM with ST-BAVA. (a) Our model takes a sequence of video frames and audio streams as input and predicts the mask of the sound sources for each video frame. (b) ST-BAVA module bidirectionally updates the image and audio features with spatial and temporal attention in sequence. M.H. stands for the multi-head. The initial audio feature from the audio backbone is used as a positional encoding for the audio feature.

computational costs [3], ST-BAVA uses the decomposed spatial and temporal cross-modal attention to reduce the memory requirements [29]. The temporal attention weight α_{time} exploits the contextual relationship between the audio-visual features per pixel, represented as follows:

$$\alpha_{\text{time}} = \text{softmax}(\frac{VA^\top}{\sqrt{C}}) \in \mathbb{R}^{HW \times T \times T}. \tag{4.1}$$

Note that the time axis $\mathbb{R}^T$ is used as the sequence dimension and the spatial axis $\mathbb{R}^{HW}$ as the batch. To match the shape of audio and visual features, spatial average pooling and repeating the input and output of the attention are applied. In spatial attention, the score map α_{space} calculates the spatial correspondence between V and A at each time step:

$$\alpha_{\text{space}} = \text{softmax}(\frac{VA^\top}{\sqrt{C}}) \in \mathbb{R}^{T \times HW \times 1}, \tag{4.2}$$

with the spatial axis $\mathbb{R}^{HW \times 1}$ used as the sequence dimension. This sequential operation of ST-BAVA first produces the image-queried audio embedding A_{vq} and uses it to produce the audio-queried image embedding V_{aq}. Repeated L times, the ST-BAVA blocks aggregate the bidirectional audio-visual relationship across the spatio-temporal dimensions to support the subsequent decoding process.

Fig. 4.2 shows the effect of temporal attention in the proposed ST-BAVA on leveraging contextual information across multiple frames. This leads to more accurate predictions of the sound source among visual candidates (violin in the left video) or judging the silent frame with no prediction (second frame of the right video) than the results without temporal attention.

4.4 Adapters

Although our proposed ST-BAVA effectively adjusts the audio-visual features at the middle of the pipeline, it relies exclusively on the visual embeddings extracted from the SAM's pre-trained image encoder that operates independently of the audio data. As Liu *et al.* [2] have revealed, audio-visual feature fusion during the image encoding stage can further enhance AVS performance. To achieve this, they have proposed Adapters that inject the audio feature into the image encoder. This work introduces

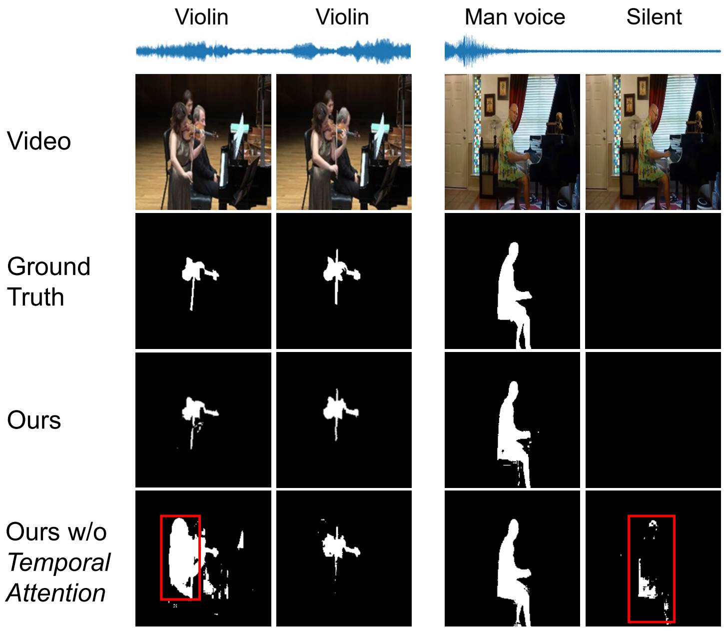


Figure 4.2: Effect of temporal attention in ST-BAVA on the audio-visual segmentation results. Our model leverages the temporal relationship across multiple frames, leading to accurate sound source predictions. Wrong prediction without temporal attention is marked in red boxes.

audio adapters to assist the ST-BAVA module in effectively fusing the audio-injected image features. Note that the Adapter does not utilize contextual or bidirectional relationships between multiple frames, as it injects audio information into the corresponding image.

For the detailed methods, in the j-th image encoder layer stage, the j-th Adapter encodes the audio embedding A from the audio backbone into $A_j^{prompt} \in \mathbb{R}^{T \times HW \times C}$ that has repeated spatial dimension. It is added to the output of the previous encoder layer E_{j-1} to be the input of j-th layer X^j, represented as:

$$X_j = E_{j-1}(X_{j-1}) + A_j^{prompt}. \tag{4.3}$$

The image embedding generated by the encoder with the Adapters proceeds through the ST-BAVA and mask decoder.

Chapter 5. Experiment

5.1 Dataset

We use the AVSBench dataset [1] designed for the audio-visual segmentation. It contains two video subsets: Single Sound Source Segmentation (S4) includes 4,932 videos, and Multi Sound Source Segmentation (MS3) contains 424 videos. In S4, a single sound source consistently appears in each video, whereas multiple sources can appear or vanish as frames progress in MS3. The videos cover 23 categories, including human voice, playing instruments, etc. Each 5-second video is split into five image frames captured per second and five audio segments, each lasting 1 second.

5.2 Evaluation Metrics

We use the mean Intersection over Union (mIoU) and F-score $\mathcal{M}_{\mathcal{F}}$ for the evaluation metrics following [1]. The mIoU computes the mean IoU between the predicted mask and the ground truth of the five frames. Note that mIoU is also known as Jaccard index $\mathcal{M}_{\mathcal{J}}$ in related works [1, 25, 23]. The F-score $\mathcal{M}_{\mathcal{F}}$ is calculated with the precision and recall, represented as $\mathcal{M}_{\mathcal{F}} = \frac{(1+\beta^2)\times precision \times recall}{\beta^2 \times precision + recall}$. β^2 is set to 0.3 in our experiments.

5.3 Implementation Details

We use pre-trained ViT-H SAM [4] for weight initialization of the SAM. The resolution of input video frames is resized to 1024×1024. We use AudioSet [30] pre-trained VGGish [31] as an audio encoder. We set the frame length $T = 5$, following [1]. During training, the parameters of the SAM's image encoder and the audio backbone encoder are not updated. We set the depth of ST-BAVA layers L to 5 in S4 and 7 in MS3. We set the attention order in ST-BAVA based on slightly better performance, but there was no significant difference. The model is trained for 250 epochs, where Adam is used as an optimizer with a learning rate of $1e-4$. Adapters that consist of two-layer MLP are inserted into all 32 image encoder layers of SAM, following the prior work [2]. We train our model with Adapters for 15 epochs from separately pre-trained ST-BAVA and Adapters.

Chapter 6. Result

Backbone	Methods	S4		MS3	
		mIoU	F-score	mIoU	F-score
PVTv2	TPAVI [1]	78.74	0.879	54.0	0.645
	AVSC [25]	81.29	0.886	59.5	0.657
	CATR [32]	81.4	0.896	59.00	0.700
	AQFormer [23]	81.6	0.894	61.1	0.721
	AVSegFormer [24]	82.06	0.899	58.36	0.693
	ECMVAE [21]	81.74	0.901	57.84	0.708
	LDM [22]	81.38	0.902	58.18	0.709
SAM	GAVS [3]	80.06	0.902	63.70	0.774
	SAMA-AVS [2]	81.53	0.886	63.14	0.691
	ST-BAVA (Ours)	**82.46**	**0.906**	**69.01**	**0.776**

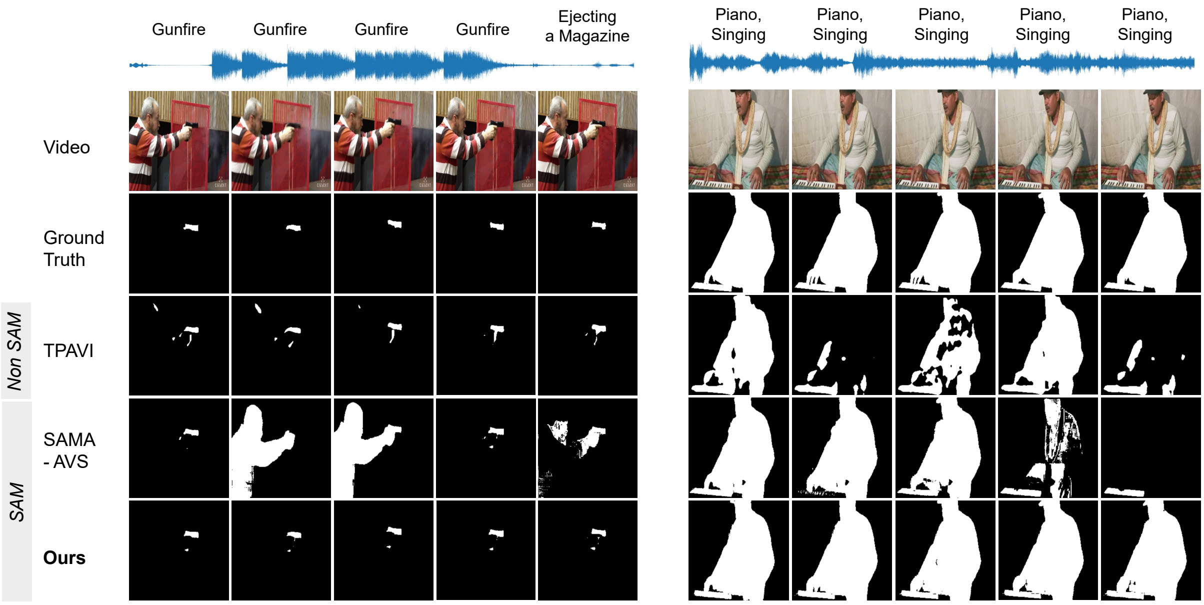
Table 6.1: Quantitative comparison with other AVS methods on the Single-source (S4) and Multi-source (MS3) subset of AVSBench dataset [1] regarding mIoU and F-score.

6.1 Quantitative Comparison

Quantitative comparisons between our method and other state-of-the-art works on the AVS benchmark are presented in Table 6.1. Our proposed ST-BAVA with SAM outperforms the state-of-the-art methods on the AVSbench in all settings. It achieves a significant performance gap in MS3, with 5.31 (8.3%) mIoU improvement compared to the previous best score (GAVS [3]). As MS3 presents the challenging task of distinguishing multiple objects corresponding to the sound within the image frame, achieving high performance on the MS3 indicates that the proposed model accurately identifies and localizes sound sources in the complex scene. We infer that our ST-BAVA supports SAM in leveraging the audio-visual correspondence between multiple sources aggregated in spatial and temporal dimensions. Note that the trainable parameters in our model are small (<4% of SAM), verifying its efficiency.

6.2 Qualitative Examples

Fig. 6.1 illustrates qualitative examples of the existing methods and ours. Our model provides examples of accurately identifying the correct sound sources among visual candidate objects. In the left video of the figure, our model successfully localizes the small sound source, the gun. In contrast, SAMA-AVS [2], another SAM-based approach without leveraging the temporal audio-visual relationship, incorrectly predicts the silent person as a sound source in several frames. It supports our claim that the proposed approach benefits from utilizing temporal relationships to comprehend the scene with more contextual information. It leads to the accurate distinction of the sounding objects among visual candidates. Moreover, our model precisely delineates the gun's refined boundaries, including the bottom part of the gun. Another non-SAM-based approach (TPAVI [1]) mispredicts the hand holding the gun



(a) Single sound source segmentation (S4) (b) Multi sound sources segmentation (MS3)

Figure 6.1: Qualitative comparison with existing methods. Our method accurately identifies sound sources across multiple frames and describes detailed object shapes, achieving the most accurate segmentation performance.

or unrelated backgrounds. In more challenging scenarios with multiple objects for the right video, our model correctly identifies the piano and the man as the sound sources and precisely portrays their shapes.

6.3 Comparison to SAM with Visual Prompts

To verify the effectiveness of the proposed audio processing approach, we compare our model with the zero-shot performance of vanilla SAM using visual prompts on the AVS benchmark in Table 6.2. Since SAM does not inherently support audio prompts, we extract the points and boxes from the ground truth mask M to guide the region of sound sources. It is an alternative to user prompts employed in zero-shot approaches with SAM in various studies [8, 11]. In the case of points, we extract the largest regions corresponding to the sound sources and select the center of mass (or random if the center doesn't lie on the object) point per region. For the boxes, the minimum external bounding boxes containing the largest contour region of sound sources in M are used.

Results in Table 6.2 show that our model performs comparable to or even better than the zero-shot performance of SAM using the ground truth regions of sound sources as a visual prompt. It verifies our method's effectiveness in handling AVS by adapting SAM with the direct capability to process audio without relying on manual prompts such as points and boxes. Note that our approach leverages the temporal cross-modal context across the multi-frame, while the vanilla SAM with visual prompts does not.

Approach	Methods	S4		MS3	
		mIoU	F-score	mIoU	F-score
Visual Prompts from G.T. mask without Training	1 Point	42.45	0.637	33.01	0.523
	3 Points	68.57	0.839	54.51	0.701
	1 Box	76.01	0.867	63.46	0.713
	3 Boxes	76.75	0.874	66.76	**0.813**
Audio Prompts with Training	w/o fusion module [2]	81.53	0.886	63.14	0.691
	+ TPAVI [1]	81.68	0.902	64.78	0.749
	+ HAN [33]	80.56	0.896	64.14	0.739
	+ CMRAN [34]	81.46	0.899	65.09	0.747
	+ JCA [35]	81.99	0.903	65.44	0.751
	+ ST-BAVA (Ours)	**82.46**	**0.906**	**69.01**	0.776

Table 6.2: Comparison of different approaches to handle the AVS with SAM. (Top) SAM receives point or box prompts guiding the region of the sound sources extracted from the ground-truth (G.T.) mask. (Bottom) For training SAM, intermediate feature fusion modules are inserted into SAM.

6.4 Ablation on Feature Fusion Modules with SAM

To prove the effectiveness of the ST-BAVA in adapting the SAM to AVS, we conducted an ablation study on the intermediate feature fusion module. For comparison, we adopt the audio-visual feature fusion modules proposed in audio-visual segmentation (Temporal Pixel-wise Audio-Visual Interaction [1, 21, 22]), audio-visual video parsing (Hybrid Attention Network [33]), audio-visual event localization (Cross-Modal Relation-Aware Networks [34]), and dimensional emotion recognition (Joint Cross-Attention [35]).

Results in Table 6.2 show that the proposed ST-BAVA outperforms other methods with SAM in the AVS benchmark. From the SAM without intermediate fusion module [2], the performance improvement by introducing the intermediate fusion module demonstrates its significance in supporting the SAM's decoder to learn complex audio-visual correspondences. However, all other fusion modules perform less than the ST-BAVA. In the case of TPAVI [1], the single integrated spatio-temporal attention is susceptible to implicit and redundant representations [29, 23], whereas the ST-BAVA separately operates spatial and temporal attention. Furthermore, bidirectionally updating audio-visual features adopted by ST-BAVA enhances their subsequent cross-attention in the SAM's mask decoder. Other modules, HAN [33], CMRAN [34], and JCA [35], struggle with adapting SAM into the AVS, as they are not designed to consider the spatial visual features that are not essential to solving their tasks.

6.5 Analysis on Attention Maps

To qualify the effect of the proposed ST-BAVA on the cross-modal features, we visualize the spatial attention score map between the audio-visual features in the middle of our model pipeline. Since the mask decoder uses the attention map to get a final prediction mask, the intermediate maps present valuable cues affecting the model's output.

In Fig. 6.2, the attention maps before the ST-BAVA module do not include any information related to the sound sources, simply depicting the boundaries of objects on the image by the pre-knowledge of the backbones. In contrast, the attention map after ST-BAVA clearly shows the high values in the sources' location, while the values in the backgrounds and the silent objects are low. It leads to the

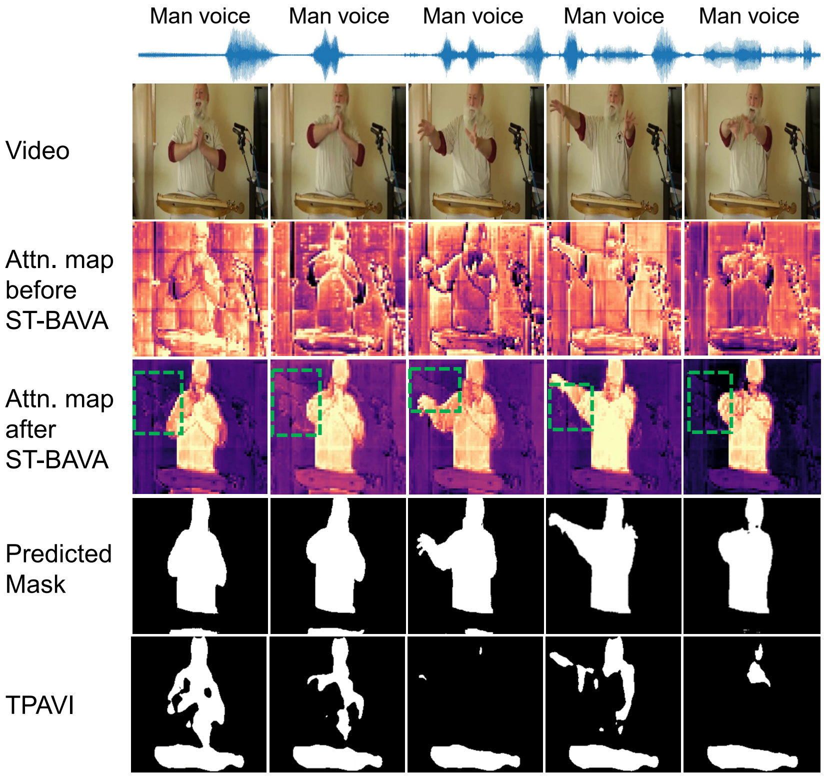


Figure 6.2: Spatial attention maps of the audio and visual embedding in the middle of our model pipeline. The attention map before ST-BAVA is calculated using features extracted from the backbones. After the ST-BAVA, the map separately represents the region of sound sources within the frames, which leads to the correct segmentation of the sources in the predicted mask. Green-boxed regions show the visual information aggregated from other frames by temporal attention (the man with multiple arms).

accurate segmentation results observed in the predicted masks, showing the effectiveness of the ST-BAVA module in judging the pixel-level audio-visual correspondence. Interestingly, after ST-BAVA, each map aggregates the visual information of other frames by temporal attention. In the green-boxed region, the man on the map has multiple arms appearing in other frames, which does not disrupt the precise prediction masks at each time step.

6.6 Ablation on Model Components

We conduct an ablation study to investigate the effectiveness of the proposed components in Table 6.3. The baseline uses spatial and unidirectional audio-to-image attention in ST-BAVA without Adapter [2]. All of the proposed components yield performance improvements in both subsets in the AVS benchmark, highlighting the effectiveness of the ST-BAVA. Notably, our proposed model without the Adapter performs well with no training or prompt-tuning of the SAM's image encoder. Utilizing the Adapter helps improve cross-modal interaction in ST-BAVA with the audio-adapted image feature, further enhancing the AVS performance.

Also, we report the performance change with the varying depth of ST-BAVA layers in Fig. 6.3. Our model performs best at the depth of 5 in the S4 subset and 7 in the MS3. It's obvious that within the MS3, there is an evident enhancement in performance beyond a depth of 6. In contrast, the S4 exhibits

Methods	S4		MS3	
	mIoU	F-score	mIoU	F-score
Baseline (Spatial A2V Attn.)	76.65	0.857	61.54	0.703
w/o Temporal Attn.	80.72	0.892	65.37	0.752
w/o Bidirectional Attn.	80.09	0.887	65.17	0.749
w/o Adapter	80.02	0.888	66.06	0.743
Full	**82.46**	**0.906**	**69.01**	**0.776**

Table 6.3: Ablation on the components of our methods.

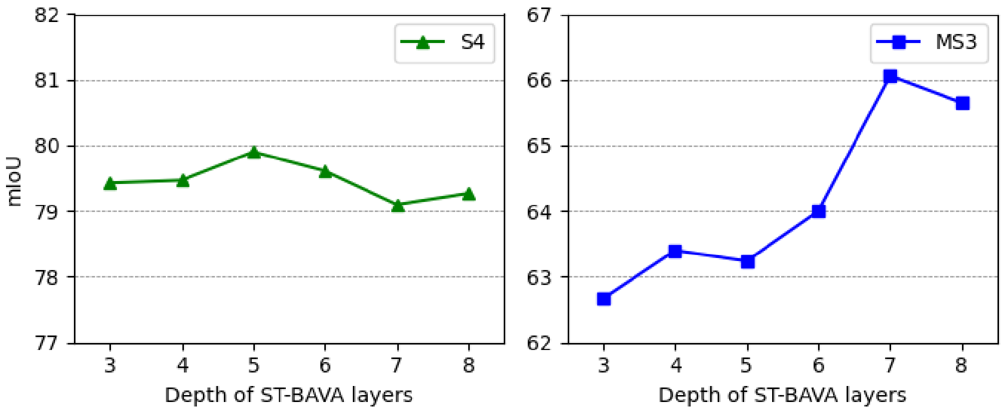


Figure 6.3: Ablation study on the depth of ST-BAVA module with mIoU of S4 and MS3 subset.

relatively consistent performance across varying depths.

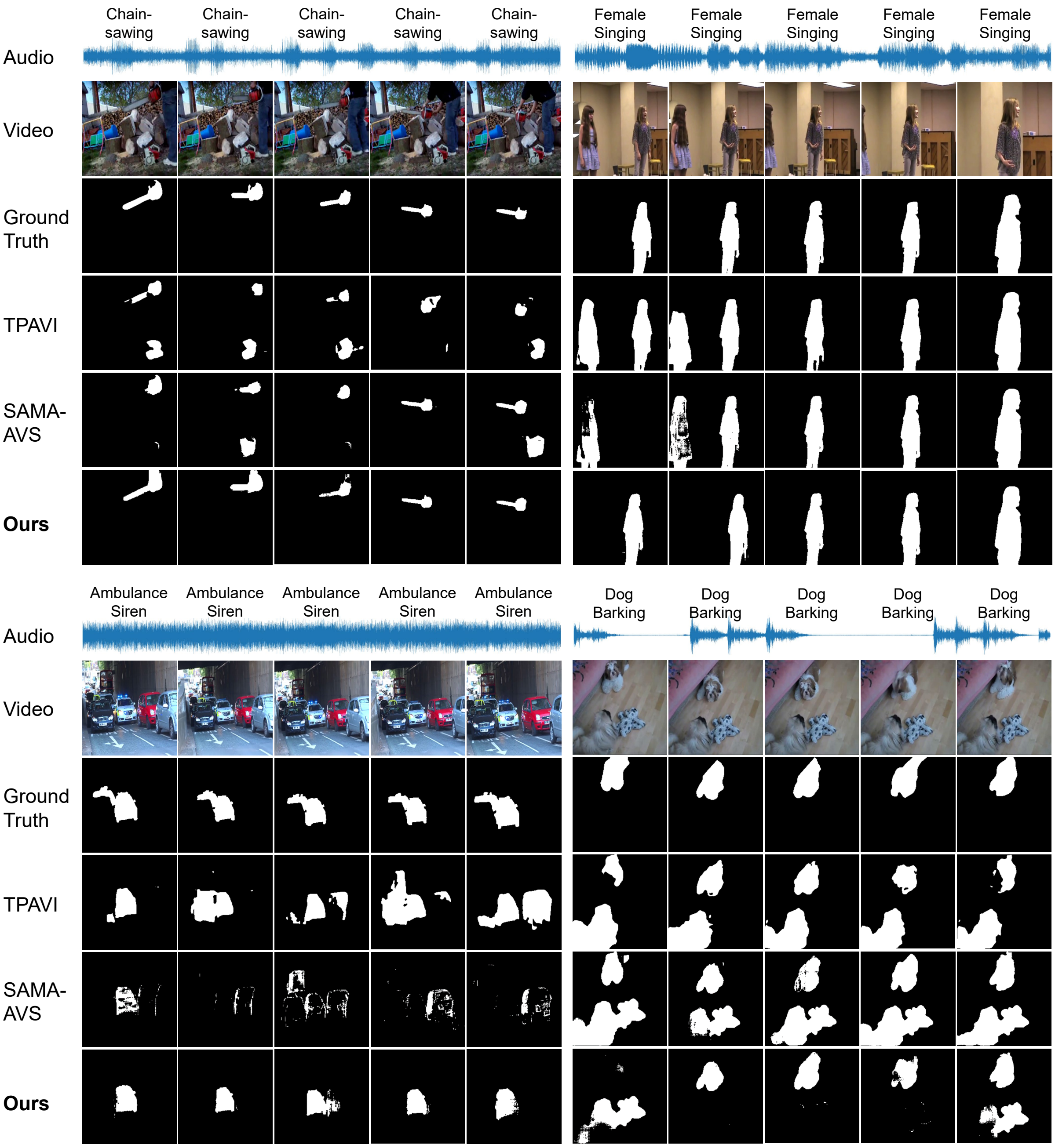


Figure 7.1: Qualitative examples where semantically similar objects appear.

7.1 Semantically Similar Object Cases

We provide examples in the scenario where semantically similar objects appear in Fig. 7.1. In the top-left videos, there are two chainsaws, but only one on the human hand emits sound. Other models are distracted by the silent chainsaw lying on the ground, whereas our model selects the correct sound source. It verifies that our approach benefits from the contextual information that the man moves the sounding chainsaw across multiple frames. Similar results are shown in the right video, where only one of the two females is singing. Nevertheless, there are also cases where our model shows insufficient results. In the bottom-left video, the prediction of our model is distracted by the adjacent cars in several frames. Moreover, in the bottom-right video, where the available visual cue is not easily noticeable, our model mispredicts the silent dog as a source in several frames. Further semantic modeling considering the instance-wise relationship could be one option for handling these cases.

7.2 Future work

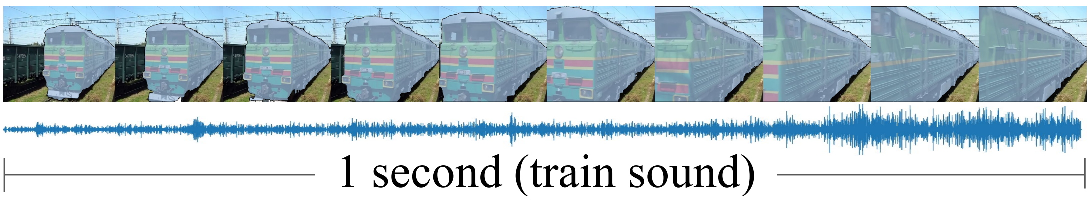


Figure 7.2: Our model's AVS prediction on the 10 FPS video in a novel category (train) that doesn't appear in AVSBench.

As SAM [4] stands as a foundation model for various image segmentation tasks [5, 6, 7, 8, 9, 10, 11, 12, 13], the generalization power of the proposed SAM with ST-BAVA can be investigated in generalized settings in terms of the sound source category, or different frame rates. For example, we test our model with a video in the evaluation set of AudioSet [30] (Fig. 7.2). The video is not in the AVSBench's 23 categories and has 10 times higher FPS than the AVSBench. The result in Fig. 7.2 highlights the potential of our model on the generalization performance regarding to the novel categories and a higher frame rate.

Chapter 8. Conclusion

We have proposed the extended Segment Anything Model (SAM) to address the Audio-Visual Segmentation (AVS) task in videos with sound. Our approach includes the Spatio-Temporal, Bidirectional Audio-Visual Attention (ST-BAVA) module, designed to analyze the spatio-temporal dynamics between multiple image and audio frames. This module, positioned between SAM's image encoder and mask decoder, enhances SAM's ability to process audio-visual information across both spatial and temporal dimensions. Within ST-BAVA, spatial attention identifies pixel-level audio-visual correlations for each video frame, while temporal attention explores the cross-modal relationships over successive frames. Our extensive experimental evaluations demonstrate that our model achieves meaningful performance on AVS benchmarks, outperforming existing state-of-the-art methods.

Bibliography

[1] J. Zhou, J. Wang, J. Zhang, W. Sun, J. Zhang, S. Birchfield, D. Guo, L. Kong, M. Wang, and Y. Zhong, "Audio–visual segmentation," in *ECCV*. Springer, 2022, pp. 386–403.

[2] J. Liu, Y. Wang, C. Ju, C. Ma, Y. Zhang, et al., "Annotation-free audio-visual segmentation," *Proc. WACV*, 2024.

[3] Y. Wang, W. Liu, G. Li, J. Ding, D. Hu, and X. Li, "Prompting segmentation with sound is generalizable audio-visual source localizer," *arXiv preprint arXiv:2309.07929*, 2023.

[4] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W. Lo, et al., "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023.

[5] J. Cen, Z. Zhou, J. Fang, W. Shen, L. Xie, X. Zhang, and Q. Tian, "Segment anything in 3d with nerfs," *arXiv preprint arXiv:2304.12308*, 2023.

[6] Q. Shen, X. Yang, and X. Wang, "Anything-3d: Towards single-view anything reconstruction in the wild," *arXiv preprint arXiv:2304.10261*, 2023.

[7] J. Wu, R. Fu, H. Fang, Y. Liu, Z. Wang, Y. Xu, Y. Jin, and T. Arbel, "Medical sam adapter: Adapting segment anything model for medical image segmentation," *arXiv preprint arXiv:2304.12620*, 2023.

[8] M. A Mazurowski, H. Dong, H. Gu, J. Yang, N. Konz, and Y. Zhang, "Segment anything model for medical image analysis: an experimental study," *Medical Image Analysis*, vol. 89, pp. 102918, 2023.

[9] Y. Wang, W. Zhou, Y. Mao, and H. Li, "Detect any shadow: Segment anything for video shadow detection," *arXiv preprint arXiv:2305.16698*, 2023.

[10] X. Zhang, C. Gu, and S. Zhu, "Sam-helps-shadow: When segment anything model meet shadow removal," *arXiv e-prints*, pp. arXiv–2306, 2023.

[11] Y. Cheng, L. Li, Y. Xu, X. Li, Z. Yang, W. Wang, and Y. Yang, "Segment and track anything," *arXiv preprint arXiv:2305.06558*, 2023.

[12] Y. Li, J. Zhang, X. Teng, and L. Lan, "Refsam: Efficiently adapting segmenting anything model for referring video object segmentation," *arXiv preprint arXiv:2307.00997*, 2023.

[13] F. Rajič, L. Ke, Y. W. Tai, C. K. Tang, M. Danelljan, and F. Yu, "Segment anything meets point tracking," *arXiv preprint arXiv:2307.01197*, 2023.

[14] S. Mo and Y. Tian, "Av-sam: Segment anything model meets audio-visual localization and segmentation," *arXiv preprint arXiv:2305.01836*, 2023.

[15] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, et al., "Parameter-efficient transfer learning for nlp," in *ICML*. PMLR, 2019, pp. 2790–2799.

[16] Y. Aytar, C. Vondrick, and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," *NeurIPS*, vol. 29, 2016.

[17] R. Arandjelovic and A. Zisserman, "Look, listen and learn," in *Proc. ICCV*, 2017, pp. 609–617.

[18] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba, "The sound of pixels," in *Proc. ECCV*, 2018, pp. 570–586.

[19] H. Chen, W. Xie, T. Afouras, A. Nagrani, A. Vedaldi, and A. Zisserman, "Localizing visual sounds the hard way," in *Proc. CVPR*, 2021, pp. 16867–16876.

[20] S. Mo and Y. Tian, "Audio-visual grouping network for sound localization from mixtures," in *Proc. CVPR*, 2023, pp. 10565–10574.

[21] Y. Mao, J. Zhang, M. Xiang, Y. Zhong, and Y. Dai, "Multimodal variational auto-encoder based audio-visual segmentation," in *Proc. ICCV*, 2023, pp. 954–965.

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

[23] S. Huang, H. Li, Y. Wang, H. Zhu, J. Dai, J. Han, et al., "Discovering sounding objects by audio queries for audio visual segmentation," *arXiv preprint arXiv:2309.09501*, 2023.

[24] S. Gao, Z. Chen, G. Chen, W. Wang, and T. Lu, "Avsegformer: Audio-visual segmentation with transformer," *arXiv preprint arXiv:2307.01146*, 2023.

[25] C. Liu, P. P. Li, X. Qi, H. Zhang, L. Li, D. Wang, and X. Yu, "Audio-visual segmentation by exploring cross-modal mutual semantics," in *Proc. ACM MM*, 2023, pp. 7590–7598.

[26] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, et al., "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," *arXiv preprint arXiv:2303.05499*, 2023.

[27] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[28] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. CVPR*, 2022, pp. 16000–16009.

[29] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?," in *ICML*, 2021, number 3, p. 4.

[30] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *ICASSP*. IEEE, 2017, pp. 776–780.

[31] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, et al., "Cnn architectures for large-scale audio classification," in *ICASSP*. IEEE, 2017, pp. 131–135.

[32] K. Li, Z. Yang, L. Chen, Y. Yang, and J. Xiao, "Catr: Combinatorial-dependence audio-queried transformer for audio-visual video segmentation," in *Proc. ACM MM*, 2023, pp. 1485–1494.

[33] Y. Tian, D. Li, and C. Xu, "Unified multisensory perception: Weakly-supervised audio-visual video parsing," in *Proc. ECCV*. Springer, 2020, pp. 436–454.

[34] H. Xu, R. Zeng, Q. Wu, M. Tan, and C. Gan, "Cross-modal relation-aware networks for audio-visual event localization," in *Proc. ACM MM*, 2020, pp. 3893–3901.

[35] R.G. Praveen, W.C. de Melo, N. Ullah, H. Aslam, O. Zeeshan, T. Denorme, M. Pedersoli, A.L Koerich, S. Bacon, et al., "A joint cross-attention model for audio-visual fusion in dimensional emotion recognition," in *CVPR*, 2022, pp. 2486–2495.

Acknowledgments in Korean

WIP

Curriculum Vitae in Korean

이 름: 선 주 형

주 소: munuwazzi@gmail.com

학 력

2015. 3. – 2018. 2. 대덕고등학교

2018. 2. – 2022. 8. 한국과학기술원 전산학부 (학사)

2022. 8. – 2024. 8. 한국과학기술원 전산학부 (석사)

경 력

2023. 3. – 2023. 8. 한국과학기술원 전산학부 조교 (CS204, 이산구조)

2021. 1. – 2021. 2. 한국전자통신연구원 연수연구원

연 구 업 적

1. **Juhyeong Seon**, Woobim Im, Sebin Lee, Jumin Lee, and Sung-Eui Yoon, "Extending Segment Anything Model into Auditory and Temporal Dimensions for Audio-Visual Segmentation", *IEEE International Conference on Image Processing (ICIP)*, Oct. 2024.
2. Jumin Lee, Sebin Lee, Changho Jo, Woobim Im, **Juhyeong Seon**, and Sung-Eui Yoon, "SemCity: Semantic Scene Generation with Triplane Diffusion", *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2024.
3. Guoyuan An, **Juhyeong Seon**, Inkyu An, Yuchi Huo, and Sung-Eui Yoon, "Topological RANSAC for instance verification and retrieval without fine-tuning", *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, Nov. 2023.
4. **Juhyeong Seon**, Jaeyoon Kim, Joo Young Kim, Young Joo Lee, Hyekyung Han, and Sung-Eui Yoon, "비디오 내 음원 위치 추정 모델의 성능 향상을 위한 클래스 인지 대조 학습 기법 제안", *정보과학회 컴퓨팅의 실제 논문지 (KTCP 2023)*, vol. 29, no. 11, pp. 518-524, Nov. 2023.
5. **Juhyeong Seon**, Guoyuan An, and Sung-Eui Yoon, "배경이 이미지 유사도 학습에 미치는 영향 분석", *한국정보과학회 2022 한국컴퓨터종합학술대회 (KCC 2022)*, vol. 49, no. 01, pp. 1991-1993, Jun. 2022.