How Can Contrastive Pre-training Benefit Audio-Visual Segmentation? A Study from Supervised and Zero-shot Perspectives

Jiarui Yu *[1]
yjr@mail.ustc.edu.cn
Haoran Li *[1]
lihaoran747@126.com
Yanbin Hao †[1]
haoyanbin@hotmail.com
Jinmeng Wu[2]
Jinmeng2004910@outlook.com
Tong Xu[1]
tongxu@ustc.edu.cn
Shuo Wang[1]
shuowang.edu@gmail.com
Xiangnan He[1]
xiangnanhe@gmail.com

[1] University of Science and Technology of China
Hefei, China

[2] Wuhan Institute of Technology
Wuhan, China

Abstract

Sharing a similar spirit with the successful contrastive language-image pre-training (CLIP), audio-aware contrastive pre-training has also exhibited its powerful ability to align instances in audio retrieval and audio-guided image generation. In this paper, we aim to extend its capabilities to the pixel level to achieve audio-visual segmentation (AVS). Specifically, we explore the following question: *how can the instance-level alignment knowledge gained from contrastive pre-training benefit pixel-level audio-visual segmentation?* To address this question, we approach the problem from two perspectives in AVS: a supervised setting and a zero-shot setting. In the supervised setting, we enhance the instance-level AudioCLIP model by incorporating a pixel-wise multi-modal fusion module. This leads to a simple yet effective model AC-FPN that enables pixel-level predictions for sounding objects, following the standard AVS training fashion. On the other hand, in the zero-shot setting, we further investigate the feasibility of promoting the Segment-Anything-Model (SAM) for AVS by proposing three prompt formulizing strategies based on instance-level contrastive pre-training models. Experimental results on both subtasks demonstrate the potential of leveraging instance-level contrastive pre-training for advancing audio-visual segmentation to the pixel level. Code is available at https://github.com/Lihr747/Sam4AVS.

*Equal Contribution.

†Yanbin Hao is the corresponding author.

1 Introduction

Researchers have been investing significant efforts to align different types of data (multi-modalities) in a shared embedding space, utilizing the large-scale contrastive pre-training. A notable example is the Contrastive Image-Language Pre-training (CLIP) [21], which uses 400 million image-text pairs for training. Recently, the scope of this pre-training methodology has expanded to incorporate audio data, aiming to align audio with mainstream modalities. For instance, Wav2CLIP [28] distills alignment knowledge from a fixed CLIP image encoder, whereas AudioCLIP [8] trains the vanilla CLIP framework with an additional audio encoder. Moreover, CLAP [5] aligns audio and text by training on noisy audio-text pairs sourced from the LAION dataset [29].

These audio-aware CLIP-like models have shown an impressive capacity to align audio with text and visual modalities. The tasks these models typically perform involve instance-level processing, such as audio-text [29] or audio-visual [28] retrieval, as well as audio-guided image generation [28]. In this paper, our objective is to investigate the viability of leveraging the pre-learned knowledge of instance-level audio-visual alignment to enhance audio-visual segmentation (AVS), a task that entails pixel-level comprehension. Simply put, we want to see if we can apply the knowledge gained from aligning audio with visuals at a broader level to a task requiring a much more detailed understanding.

AVS is a task that requires making pixel-level predictions for objects that are producing sound. This research is inspired by the successful use of CLIP in text-visual segmentation tasks, such as semantic segmentation [22], referring segmentation [27], and open-vocabulary segmentation [17]. In these tasks, CLIP has shown a strong ability to precisely align semantics and pixels, prompting us to ask: *how can the instance-level alignment knowledge gained from contrastive pre-training benefit pixel-level audio-visual segmentation?* In this paper, we focus on spatially segmenting objects based on the audio information for each frame in a video to answer this question. We approach this question from two different perspectives:

- (1) In a supervised setting, we can access video frames, audio, and ground-truth masks during training. This approach is similar to the standard AVS benchmark.
- (2) In the zero-shot setting, we aim to develop a framework capable of achieving AVS without training.

We hope to shed light on the potential benefits of contrastive pre-training for AVS through these two subtasks.

In the supervised setting, we introduce a straightforward but effective model called AC-FPN. It leverages the strengths of AudioCLIP [8] and Semantic FPN [13], one of the simplest segmentation base models. Specifically, we use AudioCLIP as the backbone for processing frames (images) and corresponding audio. We then fuse the audio feature with the high-level image feature map, followed by channel-wise concatenation. After this, we train the basic Semantic FPN decoder to predict the mask using a simple binary cross entropy (BCE) loss.

We explore three strategies to prompt the Segment-Anything-Model (SAM) [14] for AVS in a zero-shot setting. These strategies utilize the zero-shot alignment and mask generation capabilities of contrastive pre-training models and SAM. Specifically, in the first strategy, we directly employ SAM to generate masks for all objects appearing in the video frame and select the masks with high similarity scores with the audio as the final mask output. In the second and third strategies, we separately provide point and box prompts for SAM to detect the sound source. Here, the audio signal is used as a query for specific prompts, namely point and box prompts. We observe that contrastive pre-training models (e.g., AudioCLIP, CLAP) can serve as selection criteria and providers of point-prompt or box-prompt for SAM.

Experiments demonstrate that using contrastive pre-training models significantly improves the performance of supervised AVS. Moreover, our exploration of a zero-shot AVS approach, which does not require training, suggests the possibility of generating an image mask for sound signals without training.

2 Related Work

2.1 Multi-modal Contrastive Pre-training

Contrastive learning [20] aims to differentiate between paired samples and unpaired samples by training the model to map paired samples to nearby points and unpaired samples to distant points in the representation space. In multi-modal scenarios, contrastive learning is applied to align distinct modalities. In practice, contrastive pre-training models use distinct encoders to extract modality features and perform in-batch contrastive learning with large-scale multi-modal pairs. This training paradigm is originally from the visual-text domain, where CLIP [21] and ALIGN [12] use simple contrastive learning to train a dual-encoder for image and visual representation with 400M and 1.8B web-collected image-text pairs. Most related to our work is a series of contrastive pre-training models considering audio signals. Video streams offer natural audio-visual pairs, making it possible to achieve visual-audio alignment by leveraging video datasets. Wav2CLIP [28] distils knowledge from CLIP by training an audio encoder under the supervision of the fixed CLIP image encoder. AudioCLIP [8] extends an audio encoder to the CLIP framework by training it with frames and textual labels that correspond to audio from Audioset [6]. Besides learning audio modality with visuals, CLAP [29] connects language and audio by conducting contrastive training on 630k audio-text pairs.

2.2 Audio-Visual Localization and Segmentation

In contrast to instance-level audio-visual matching [3], audio-visual localization and segmentation aim to ground the audio temporally or spatially in the input visual data, such as images or videos. Temporal localization tasks include audio-visual event localization[4, 24], which predicts the event of video segments with pre-defined event labels, and audio-visual video parsing[18, 25], which divides unconstrained videos into a set of video events associated with event categories. Sound source localization [2, 11], a spatial localization task, aims to locate image regions related to the sound maker, but its results are usually at the patch level, providing insufficient information about the object's actual shape. To address this issue, Zhou et al. [30] proposed a pixel-level annotated audio-visual segmentation benchmark (AVSBench) and a baseline method called TPAVI. In this study, we also focus on audio-visual segmentation and we evaluate the benefits of leveraging contrastive pre-training models on the AVSBench.

3 AC-FPN: Simple Supervised AVS with AudioCLIP

In this section, we introduce a simple yet efficient supervised audio-visual segmentation (AVS) method, AudioCLIP-FPN (AC-FPN). The purpose of AC-FPN is to transfer the instance-level audio-visual alignment knowledge from a contrastive pre-training model to the pixel-level audio-visual segmentation scenario. To assess the effectiveness of AudioCLIP knowl-

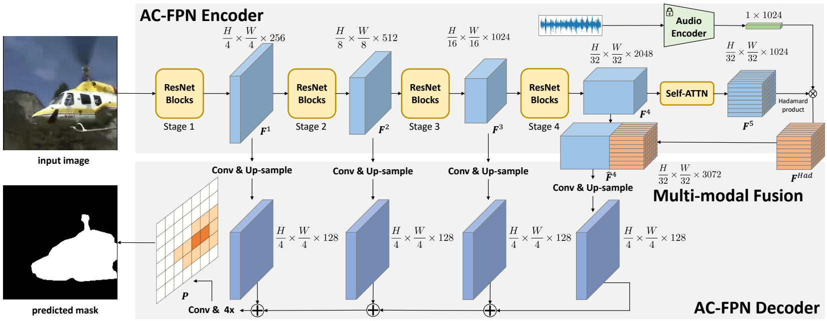

AC-FPN Encoder
input image
ResNet Blocks
Stage 1
Stage 2
Stage 3
Stage 4
$\frac{H}{4}\times\frac{W}{4}\times 256$
$\frac{H}{8}\times\frac{W}{8}\times 512$
$\frac{H}{16}\times\frac{W}{16}\times 1024$
$\frac{H}{32}\times\frac{W}{32}\times 2048$
F^1
F^2
F^3
F^4
Audio Encoder
1×1024
Self-ATTN
$\frac{H}{32}\times\frac{W}{32}\times 1024$
F^5
Hadamard product
F^{Had}
$\hat{F}^4$
$\frac{H}{32}\times\frac{W}{32}\times 3072$
Multi-modal Fusion
Conv & Up-sample
$\frac{H}{4}\times\frac{W}{4}\times 128$
P
Conv & 4x
predicted mask
AC-FPN Decoder

Figure 1: Overall framework of AC-FPN with Hadamard production fusion. AC-FPN integrates the AudioCLIP encoders into the Semantic FPN framework incorporating a multi-modal fusion module to sense the audio query.

edge, we keep our AC-FPN **as simple as possible**. As illustrated in Figure 1, AC-FPN comprises three essential modules: encoder, multi-modal fusion, and decoder.

Encoder. We adopt ESResNeXt [7] to extract the audio feature $\boldsymbol{f}^a$, and we use ResNet-50 [9] to construct the bottom-up visual feature maps $\{\boldsymbol{F}^i\}_{i=1}^4$. The two encoders are both from AudioCLIP pre-training and we keep the audio encoder fixed. Similar to CLIP, AudioCLIP appends an attention pooling layer after the $\boldsymbol{F}^4$. Specifically, AudioCLIP concatenates $\boldsymbol{F}^4$, as a series of image tokens, with its average pooling and then fed into a self-attention [26] layer to produce a new visual feature map $\boldsymbol{F}^5$ and a global visual representation $\boldsymbol{f}^v$, i.e., $[\boldsymbol{f}^v, \boldsymbol{F}^5] = \text{Self-ATTN}([\overline{\boldsymbol{F}}^4, \boldsymbol{F}^4])$. Inspired by [22], we conjecture that $\boldsymbol{F}^5$ contains sufficient semantic information as an audio-aware feature map and behaves similarly to the global feature $\boldsymbol{f}^v$ due to the token symmetry of self-attention. Therefore, we posit that fusing the audio feature $\boldsymbol{f}^a$ and the audio-aware visual feature map $\boldsymbol{F}^5$ is the key to effective AVS.

Multi-modal Fusion. As previously discussed, the $\boldsymbol{F}^5$ feature map in AudioCLIP is adept at audio-semantic sensing. Hence, we fuse the audio feature $\boldsymbol{f}^a$ into the $\boldsymbol{F}^5$ visual feature map by a pixel-wise fusion to obtain a fused feature map. We explore two straightforward fusion strategies: *Hadamard production* and *concatenation*. (1) *Hadamard production fusion* involves computing the element-wise Hadamard product between $\boldsymbol{f}^a$ and visual feature at each position in $\boldsymbol{F}^5$, as shown below:

$$\boldsymbol{F}^{\text{Had}}_{i,j} = \boldsymbol{f}^a \odot \boldsymbol{F}^5_{i,j}, \text{where } i \in [1, H/32], j \in [1, W/32], \tag{1}$$

where $\odot$ is Hadamard product. H and W represent the height and width of the original input images. (2) *Concatenation fusion* concatenates the audio feature $\boldsymbol{f}^a$ with visual embeddings at each position in $\boldsymbol{F}^5$, as follows:

$$\boldsymbol{F}^{\text{cat}}_{i,j} = [\boldsymbol{f}^a, \boldsymbol{F}^5_{i,j}], \text{where } i \in [1, H/32], j \in [1, W/32]. \tag{2}$$

To integrate fused feature map into the FPN framework, we merge the fused feature map $\boldsymbol{F}^{\text{fusion}}$ ($\boldsymbol{F}^{\text{Had}}$ or $\boldsymbol{F}^{\text{cat}}$) to the $\boldsymbol{F}^4$ feature map using channel-level concatenation, obtaining a new feature map $\hat{\boldsymbol{F}}^4 = [\boldsymbol{F}^4, \boldsymbol{F}^{\text{fusion}}]$. As a result, the bottom-up stage of FPN produces four feature maps: $[\boldsymbol{F}^1, \boldsymbol{F}^2, \boldsymbol{F}^3, \hat{\boldsymbol{F}}^4]$, which are subsequently used in the decoding stage.

Decoder and Loss. The AC-FPN decoder is designed to make a mask prediction based on the bottom-up feature maps obtained from the encoder and multi-modal fusion stages. Our decoder employs the same head and neck as the Semantic FPN [13]. The decoding stage produces a score map $\boldsymbol{P}$, and then a simple BCE loss is utilized to train the model to correctly classify each pixel.

To clarify the simpleness of AC-FPN, we compare AC-FPN to the baseline method, TPAVI-ResNet50 [30], showing four apparent differences: (1) Encoder. TPAVI utilizes ResNet-50 [9] trained on ImageNet [23] and VGGish [10] trained on Audioset as the visual and audio encoder. We take both encoders from AudioCLIP. Note that the audio encoders are fixed in AC-FPN and TPAVI. (2) Fusion. TPAVI uses a temporal pixel-wise audio-visual attention module to integrate the audio and visual information of all frames. In contrast, our AC-FPN uses a simple Hadamard product or concatenation fusion. (3) Decoder. Multi-size kernels are used in TPAVI's neck implementation to enhance the feature, whereas AC-FPN uses vanilla Semantic FPN with single-size kernels. (4) Loss. We use simple BCE loss abandoning the KL divergence for regularization used in TPAVI.

4 Exploring Zero-shot AVS with Contrastive Pre-training

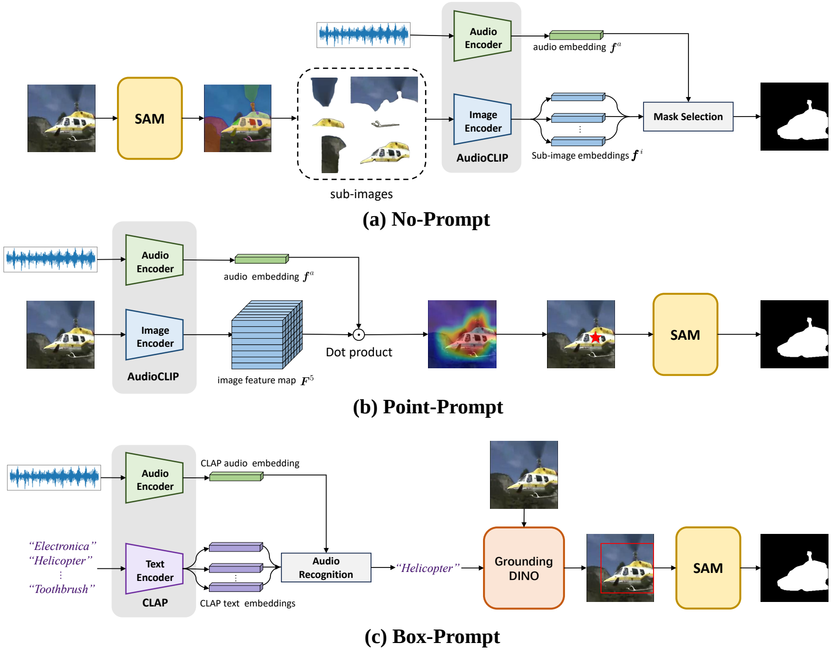
Recently, Segment Anything Model (SAM) [14] shows its accurate mask prediction in broad scenarios due to the large-scale training on the 1B-mask dataset. Additionally, SAM is capable of promptable segmentation, allowing it to generate valid segmentation masks based on pre-designed prompts. However, SAM's prompts are currently limited to *box*, *point*, and *mask*, which renders it incapable of responding to auditory prompts. In contrast, large-scale audio-aware contrastive pre-training (ACP) aligns audio signals with other modalities, leading to zero-shot audio understanding, but fails in pixel-level prediction. In this section, we explore several approaches to leverage the power of zero-shot mask prediction from SAM and zero-shot audio understanding from ACP models to achieve zero-shot audio-visual segmentation without any training.

Our core idea is to create an *interface* between visual-audio signals and SAM, by converting visual and audio features into suitable prompts. Since the current SAM does not seem to support a single mask as prompt [1], we investigate three strategies: No-Prompt, Point-Prompt and Box-Prompt. As shown in Figure 2, the No-Prompt strategy generates all possible masks using SAM and then selects related ones, while, the Point-Prompt and Box-Prompt strategies prompt the SAM using points and boxes, respectively. It is worth noting that these three strategies do not require any training.

No-Prompt. According to SAM, it can automatically generate masks for all objects, without any prompting. The intuitive solution provided by SAM is to segment all objects in the image content and then rank their masks based on the audio query. In particular, we crop the image according to each mask and apply padding to create individual sub-images. We employ AudioCLIP to encode these sub-images and the given audio separately. Then, we calculate the cosine similarities between the sub-image embeddings and the audio embedding. Subsequently, we select all sub-image embeddings whose cosine similarities exceed a certain threshold and concatenate their original masks to obtain the final segmentation result.

Point-Prompt. As suggested by [15, 16], the last feature map of CLIP image encoder can provide valuable semantic information for cross-modal explanations. In the Point-Prompt strategy, we aim to obtain positive and negative points from the cross-modal heatmap to prompt SAM for mask prediction. To achieve this, we leverage AudioCLIP, an advanced visual-audio alignment model. Specifically, we first generate a heatmap $\boldsymbol{H}$ by calculating the

[1] https://github.com/facebookresearch/segment-anything/issues/169



(a) No-Prompt

(b) Point-Prompt

(c) Box-Prompt

Figure 2: Frameworks of three zero-shot strategies: No-Prompt, Point-Prompt and Box-Prompt. No-Prompt uses SAM to segment anything and selects related ones. Point-Prompt and Box-Prompt mine points or boxes from contrastive pre-training models to prompt SAM.

cosine similarity of the last-stage audio feature $\boldsymbol{f}^a$ and image feature $\boldsymbol{F}^5$ at each position. Heatmap value at position (i,j) can be represented as $\boldsymbol{H}_{i,j} = \cos(\boldsymbol{f}^a, \boldsymbol{F}^5_{i,j})$.

However, we observed the heatmap $\boldsymbol{H}$ is against human understanding, i.e., the region more related to the audio in the heatmap has a lower score. Li et al. [15] revealed a similar phenomenon in the vanilla CLIP due to attentive pooling. We adopt CLIP architecture surgery [16] and Reverse [15] to further process the heatmap, leading to a satisfactory heatmap. Afterwards, we use the min-max normalization technique to rescale the heatmap. Leveraging the heatmap that roughly reflects the regions related to a given audio, we extract points from the heatmap using the following three methods: (1) **Global** method, where we choose the point with the highest score as the positive point, and the one with the lowest score as the negative point. (2) **Local** method, where we select the peaks and the valleys of the feature map as positive and negative points, respectively. (3) **Dense** method, where we set points with scores higher than a given threshold as positive and select an equal number of low-scoring points as negative points, inspired by [16]. Prompted by points extracted from the heatmap, SAM can convert the coarse mask in the heatmap to a more accurate mask.

Box-Prompt. To produce boxes prompting SAM, we rely on the existing tool Grounded SAM [1], which combines Grounding DINO [19] and SAM to segment objects in the open vocabulary. However, as Grounded SAM only takes text as input, transforming auditory signals into text presents a challenge. Fortunately, the CLAP [29] model has shown to be effective at classifying audio. Therefore, we employ CLAP to predict the category to which the audio belongs. We select class name with top-1 score for single-source audio, while we select top-2 class names for multi-source audio. To maintain the zero-shot setting strictly, we do not use category names in the AVSBench and instead adopt the vocabulary of AudioSet, which has 527 category names. Using the predicted categories, we query Grounding DINO to obtain the box predictions, which are then used to query SAM for mask prediction.

5 Experiments

In this section, we first introduce the dataset, settings, evaluation metrics and baselines, and then make a thorough examination, including main results, ablation study on different model components, heatmap visualization and case study.

Dataset. We conduct our experiments on the Audio-Visual Benchmark (AVSBench) [30], which includes 5-second video clips from YouTube paired with corresponding audio signals. Each clip is represented by five frames. The AVSBench consists of two subsets: Single Sound Source Segmentation (S4) and Multiple Sound Source Segmentation (MS3). The S4 subset includes videos featuring only one sounding object, while the MS3 subset contains videos with two or more sounding objects. The train/validation/test split in S4 and MS3 is 3,452/740/740 and 296/64/64, respectively. In the MS3 subset, each video frame has a binary mask annotation, while in S4 only the first frame of each video is annotated.

Settings. We conduct experiments in two distinct settings: supervised and zero-shot. The supervised setting uses ground-truth mask labels for training. While the zero-shot setting leverages existing contrastive pre-training models and advanced segmenter SAM to develop an AVS system without further training. We evaluate our model on both S4 and MS3 subsets.

Evaluation metrics and baselines. Following AVSBench, we choose Mean Intersection over Union (mIoU) and F-score[2] as the evaluation metrics. For the supervised setting, we choose TPAVI as the baseline, which is provided by AVSBench. As for the zero-shot setting, to the best of our knowledge, no methods with the same setting have been proposed before. Therefore, we design two simple baselines for the zero-shot scenario: (1) Random-SAM. Select a random point to prompt SAM. (2) Full-mask. All pixels are predicted as corresponding to the audio.

5.1 Main Results

In the supervised setting, we compare our proposed AC-FPN with TPAVI-ResNet50. To ensure a fair comparison, we set the batch size to 4 and froze the audio encoder, which is similar to TPAVI. In the zero-shot setting, we test three prompting strategies and compare them to Random-SAM and Full-mask.

Method	S4		MS3		Fixed Params. ↓	Tunable Params. ↓
	mIoU ↑	F-score ↑	mIoU ↑	F-score ↑		
TPAVI-ResNet50 [30]	72.79	.848	47.88	.578	72.1M	91.4M
AC-FPN (Hadamard)	77.12	.874	**49.95**	.635	**32.1M**	**68.0M**
AC-FPN (Concatenation)	**77.29**	**.879**	48.63	**.637**	**32.1M**	68.2M

Table 1: Performance comparison on AVSBench test split in the supervised setting.

Results in the supervised setting. As shown in Table 1, AC-FPN outperforms TPAVI baseline significantly with fewer fixed (audio encoder) and tunable parameters. Though AC-FPN does not integrate information from other frames and uses lightweight encoder-decoder,

[2] Following AVSbench, F-score is set as: $F_\beta = \frac{(1+\beta^2)\times \text{precision} \times \text{recall}}{\beta^2 \times \text{precision} + \text{recall}}$, where $\beta^2 = 0.3$

simple fusion operation and loss, AC-FPN (Hadamard) gains +4.3%(S4) and +2.0%(MS3) mIoU improvement over TPAVI leveraging the contrastive pre-training backbone and its inherent alignment knowledge.

Method	S4		MS3	
	mIoU	F-score	mIoU	F-score
Random-SAM	7.0	.240	11.5	.187
Full-mask	19.0	.226	12.7	.170
No-Prompt	23.8	.358	19.7	.242
Point-Prompt(global)	27.2	.424	19.4	.279
Point-Prompt(local)	30.7	.416	20.0	.270
Point-Prompt(dense)	40.3	.515	28.8	.333
Box-Prompt	**51.2**	**.615**	**41.8**	**.478**

Table 2: Performance comparison on AVSBench test split in the zero-shot setting).

Method	S4	
	mIoU	F-score
Hadamard	77.12	.874
Concatenation	77.29	.879
Scoremap	76.67	.874
Audio-only	76.33	.873

Table 3: Performance comparison with different multi-modal fusions in the supervised AVS (test split).

Results in the zero-shot setting. We evaluate No-Prompt, Point-Prompt, and Box-Prompt methods, and compare their performance with Random-SAM and Full-mask baselines on AVS with zero-shot setting, as shown in Table 2. Prompt-based SAM methods outperform Random-SAM and Full-mask methods, indicating contrastive pre-training models provide effective prior knowledge for audio-aware pixel-level understanding. No-Prompt selects the most related sub-images, but it is less effective due to its poor perception of low-resolution sub-images and incorrect selection of complementary backgrounds. Point-Prompt achieves better performance than No-Prompt, using the heatmap from AudioCLIP. The dense version of Point-Prompt outperforms local and global methods because more points lead to more robust prompting. Box-Prompt bridges the audio signal and detection model using the category name and provides boxes as prompts, achieving the best performance.

5.2 Ablation Study

This section presents an ablation study to investigate the effect of different multi-modal fusion approaches, variant operations for heatmap generation of Point-Prompt and changing the category name list of Box-Prompt.

Visual-Enc.	Pre-train	Audio-Enc.	Pre-train	S4_mIoU	MS3_mIoU
R50	Contrastive	ESResNeXt	Contrastive	77.12	49.95
R50	Contrastive	ESResNeXt	AudioSet	76.89	49.20
R50	ImageNet	ESResNeXt	Contrastive	67.88	37.91

Table 4: AC-FPN performance on AVSBench test split with different pre-training tasks.

Method	S4	
	mIoU	F-score
CLIP-surgery	3.2	.227
Reverse	25.4	.414
CLIP-surgery + Reverse	**40.3**	**.515**

Table 5: Point-Prompt performance with different heatmap generation operations on AVSBench test set (zero-shot).

Category list	S4		MS3	
	mIoU	F-score	mIoU	F-score
Audioset	51.2	.615	**41.8**	**.478**
AVSBench	**57.6**	**.678**	40.3	.465

Table 6: Performance change of Box-Prompt with different category name list on AVS-Bench test split (zero-shot).

Recognition Model	S4		MS3	
	mIoU	F-score	mIoU	F-score
CLAP	**51.2**	**.615**	**41.8**	**.478**
AudioCLIP	41.1	.516	32.7	.398

Table 7: Box-Prompt performance on AVSBench test split (zero-shot) with different audio recognition models.

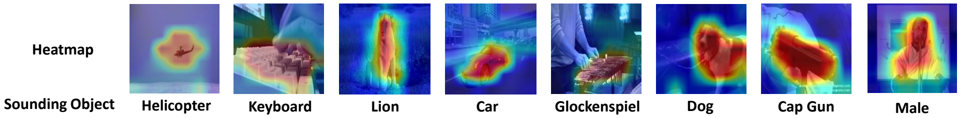


Figure 3: Visualization for heatmap adopted in Point-Prompt.

Multi-modal fusion (in supervised AC-FPN). To identify the key of multi-modal fusion in AC-FPN, we test other fusion approaches, i.e., $\boldsymbol{F}^{\text{fusion}}$ computing, as shown in Table 3. The Scoremap approach sets the 1-channel cosine score map of $\boldsymbol{f}^a$ and $\boldsymbol{F}^5$ as $\boldsymbol{F}^{\text{fusion}}$. The result is not good because a single channel cannot represent rich prior alignment knowledge. The Audio-only approach simply repeats $\boldsymbol{f}^a$ as $\boldsymbol{F}^{\text{fusion}}$, working even worse than the Scoremap fusion indicating the combination of $\boldsymbol{f}^a$ and high-level feature map $\boldsymbol{F}^5$ is the key to fusion.

Contrastive pre-training (in supervised AC-FPN). To underscore the significance of contrastive pre-training, we initialize the visual and audio encoders with alternative pre-training tasks. As demonstrated in Table 4, performance degrades when employing alternative pre-training tasks for the audio encoder (e.g., AudioSet) or the visual encoder (e.g., ImageNet). These findings highlight the importance of pre-learned knowledge in aligning visual and audio elements through contrastive pre-training.

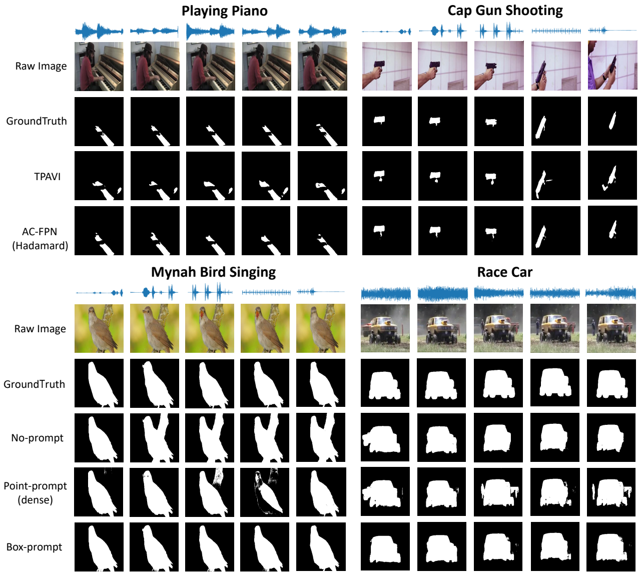
Heatmap generation of Point-Prompt. For the generation of an AudioCLIP-based audio-visual heatmap, two techniques, CLIP-surgery and Reverse are used. We show the results of Point-Prompt (dense) with single CLIP-surgery or Reverse in Table 5. The comparison suggests the two operations are both crucial to the heatmap generations.

Category list of Box-Prompt. In the Box-Prompt, we adopt a third-party category name list from Audioset, containing 527 categories. To understand how the Box-Prompt relies on the category names, we replace the Audioset list with the origin category list (23 categories) from AVSBench, shown in Table 6. Compared to Box-Prompt using Audioset list, the one with AVSBench list only performs better by a small margin on S4, and even has a lower MS3 score, which indicates Box-Prompt is not sensitive to the category name list.

Audio recognition model of Box-Prompt. In Box-Prompt, we employ CLAP, an audio-language model, to recognize the audio's category as the textual prompt for Grounded SAM. Additionally, AudioCLIP, as a substitution for CLAP, can also transcribe auditory signals into text. We test Box-Prompt's performance using both models, shown in Table 7. CLAP exhibits superior performance compared to AudioCLIP, due to its enhanced generalizability.

5.3 Heatmap Visualization and Case Study

First, we show the Point-Prompt's heatmap in Figure 3. Though it shows good activation for the sounding objects, we note that the activation area can sometimes be much larger



(a) Case study for our proposed AC-FPN (Hadamard) in the supervised setting. AC-FPN predicts more accurate masks compared to TPAVI.

(b) Case study for three prompting strategies in the zero-shot setting. No-prompt sometimes mixes objects and backgrounds; Point-Prompt activates wrong region at some frames; Box-Prompt can predict robust and satisfactory results.

Figure 4: Case study for results of our proposed AC-FPN (Hadamard) in the supervised AVS setting and three prompting strategies in the zero-shot AVS setting in the S4 subset.

than the object (e.g. the *helicopter*), leading to incorrect masks. We plan to eliminate these incorrect predictions in future work. In Figure 4(a), our proposed AC-FPN exhibits more accurate mask prediction in the supervised setting, distinguishing objects such as *piano keys* and *gun* from *hands*. We believe that AC-FPN's ability to understand spatial semantics is facilitated by the rich semantic alignment knowledge obtained from contrastive pre-training models. In Figure 4(b), both Point-Prompt and Box-Prompt yield good results. However, the performance of Point-Prompt is sometimes suboptimal, due to unstable heatmap activation. In contrast, Box-Prompt's robustness is evident, owing to utilizing existing powerful models.

6 Conclusion

We have presented a novel audio-visual segmentation (AVS) pipeline that harnesses the potential of instance-level contrastive pre-training to advance pixel-level AVS. Our approach encompasses two perspectives of AVS settings: a supervised setting and a zero-shot setting. Through comprehensive experimentation, we showcase a range of strategies that effectively utilize instance-level alignment knowledge to attain pixel-level AVS. The experimental results validate the efficiency and effectiveness of the proposed methods, underscoring their promise and potential in the field of AVS.

7 Acknowledgement

The work was supported by the National Key Research and Development Program of China under Grants 2020YFB1406703 and by the National Natural Science Foundation of China (No. 62101524).

References

[1] Grounded segment anything. https://github.com/IDEA-Research/Grounded-Segment-Anything, 2023.

[2] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Localizing visual sounds the hard way. In *CVPR*, pages 16867–16876, 2021.

[3] Jason Cramer, Ho-Hsiang Wu, Justin Salamon, and Juan Pablo Bello. Look, listen, and learn more: Design choices for deep audio embeddings. In *ICASSP*, pages 3852–3856. IEEE, 2019.

[4] Bin Duan, Hao Tang, Wei Wang, Ziliang Zong, Guowei Yang, and Yan Yan. Audio-visual event localization via recursive fusion by joint co-attention. In *WACV*, pages 4013–4022, 2021.

[5] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In *ICASSP*, pages 1–5. IEEE, 2023.

[6] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *ICASSP*, pages 776–780. IEEE, 2017.

[7] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Esresne (x) t-fbsp: Learning robust time-frequency transformation of audio. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021.

[8] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio. In *ICASSP*, pages 976–980. IEEE, 2022.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[10] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *ICASSP*, pages 131–135. IEEE, 2017.

[11] Di Hu, Feiping Nie, and Xuelong Li. Deep multimodal clustering for unsupervised audiovisual learning. In *CVPR*, pages 9248–9257, 2019.

[12] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, pages 4904–4916. PMLR, 2021.

[13] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *CVPR*, pages 6399–6408, 2019.

[14] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[15] Yi Li, Hualiang Wang, Yiqun Duan, Hang Xu, and Xiaomeng Li. Exploring visual interpretability for contrastive language-image pre-training. *arXiv preprint arXiv:2209.07046*, 2022.

[16] Yi Li, Hualiang Wang, Yiqun Duan, and Xiaomeng Li. Clip surgery for better explainability with enhancement in open-vocabulary tasks. *arXiv preprint arXiv:2304.05653*, 2023.

[17] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. *arXiv preprint arXiv:2210.04150*, 2022.

[18] Yan-Bo Lin, Hung-Yu Tseng, Hsin-Ying Lee, Yen-Yu Lin, and Ming-Hsuan Yang. Exploring cross-video and cross-modality signals for weakly-supervised audio-visual video parsing. volume 34, pages 11449–11461, 2021.

[19] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

[20] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021.

[22] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *CVPR*, pages 18082–18091, 2022.

[23] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115:211–252, 2015.

[24] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *ECCV*, pages 247–263, 2018.

[25] Yapeng Tian, Dingzeyu Li, and Chenliang Xu. 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*, pages 436–454. Springer, 2020.

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017.

[27] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11686–11695, 2022.

[28] Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. Wav2clip: Learning robust audio representations from clip. In *ICASSP*, pages 4563–4567. IEEE, 2022.

[29] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP*, pages 1–5. IEEE, 2023.

[30] Jinxing Zhou, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, and Yiran Zhong. Audio–visual segmentation. In *ECCV*, pages 386–403. Springer, 2022.