Weakly-Supervised Audio-Visual Segmentation

Shentong Mo, Yixuan Chen, Ding Bai
MBZUAI
{shentong.mo,yixuan.chen,ding.bai}@mbzuai.ac.ae

Abstract

Audio-visual segmentation is a challenging task that aims to predict pixel-level masks for sound sources in a video. Previous work applied a comprehensive manually designed architecture with countless pixel-wise accurate masks as supervision. However, these pixel-level masks are expensive and not available in all cases. In this work, we aim to simplify the supervision as the instance-level annotation, *i.e.*, weakly-supervised audio-visual segmentation. We present a novel Weakly-Supervised Audio-Visual Segmentation framework, namely WS-AVS, that can learn multi-scale audio-visual alignment with multi-scale multiple-instance contrastive learning for single-source audio-visual segmentation Extensive experiments on AVSBench demonstrate the effectiveness of our WS-AVS in the weakly-supervised audio-visual segmentation over previous baselines. Demo, code, and pre-trained models are available at our project website: https://wsavs.github.io/.

1 Introduction

When we hear a dog barking, we are naturally aware of where the dog is in the room due to the strong correspondence between audio signals and visual objects in the world. This human perception intelligence attracts many researchers to explore audio-visual joint learning for visual sound localization and segmentation. In this work, we aim to segment sound source masks from both frames and audio in a video without relying on expensive annotations of pixel-level masks, *i.e.*, only the instance-level annotation is given for training.

Audio-visual segmentation (AVS) [1] is a recently rising problem that predicts a pixel-level map of objects that produce sound at the time of frames. A related problem to this recent task is sound source localization, which aims to locate visual regions within the frame that correspond to the sound. Sound source localization (SSL) [2, 3, 4, 5, 6, 7, 8, 9, 10] estimates a rough location of the sounding objects at a patch level, while the goal of AVS is to estimate pixel-wise segmentation masks for all sound sources. AVS requires a more comprehensive manually designed architecture than SSL since it needs to identify the exact shape of sounding objects. However, this complicated network for AVS requires countless pixel-wise accurate masks as supervision, which costs numerous labor and resources. In contrast, we aim to get rid of these costly pixel-level annotations to achieve weakly-supervised audio-visual segmentation, as shown in Figure 1.

Many weakly-supervised methods [11, 12, 13, 14, 15, 16, 17, 18, 19, 20] have been proposed to be successful in the visual segmentation community, where they replaced accurate pixel-level annotations with weak labels, including scribbles [13, 15], bounding boxes [11, 12], points [14], and image-level class labels [16, 17]. Typically, CAM [18] was proposed with a CNN-based image classifier to generate pseudo segmentation masks at the pixel level using coarse object localization maps. With coarse localization maps, following works tried to obtain the full extent of objects by regions expansion [21, 22, 23], stochastic inference [24], boundary constraints [25, 26, 27], or object regions mining and erasing [28, 29]. However, only one modality, *i.e.*, image is involved in these aforementioned CAM-based weakly-supervised approaches. In this work, we need to tackle with two distinctive modalities (audio and vision), where a well-designed cross-modal fusion is required to aggregate audio-visual features with high similarity in the semantic space.

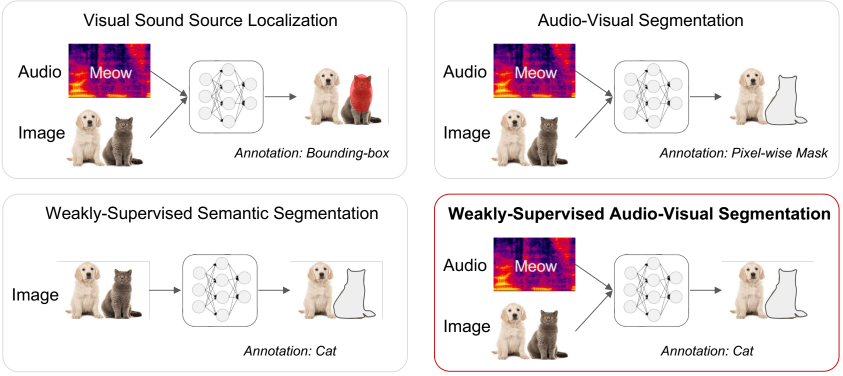
Visual Sound Source Localization
Audio
Meow
Image
Annotation: Bounding-box
Audio-Visual Segmentation
Audio
Meow
Image
Annotation: Pixel-wise Mask
Weakly-Supervised Semantic Segmentation
Image
Annotation: Cat
Weakly-Supervised Audio-Visual Segmentation
Audio
Meow
Image
Annotation: Cat

Figure 1: Comparison of visual sound source localization [2], audio-visual segmentation [1], weakly-supervised visual segmentation [31], and the proposed weakly-supervised audio-visual segmentation. *Visual Sound Source Localization* (Top Left): predicting sound source maps from both audio and image with annotations of bounding-boxes; *Audio-Visual Segmentation* (Top Right): predicting sound source masks from both audio and image with annotations of pixel-wise masks; *Weakly-Supervised Semantic Segmentation* (Bottom Left): predicting object masks from only image with instance-level annotations; **Weakly-Supervised Audio-Visual Segmentation** (Bottom Right): predicting sound source masks from both audio and image with instance-level annotations.

The main challenge is that we do not have pixel-level segmentation masks for this new weakly-supervised multi-modal problem during training. Without this supervision, the performance of AVS [1] on single-source audio-visual segmentation degrades significantly, as observed in our experiments in Section 4.2. In the meanwhile, previous weakly-supervised semantic segmentation [18, 30] and visual sound source localization [9] methods have two main challenges. First, no multi-modal constraint is applied for pixel-wise mask prediction in the previous weakly-supervised semantic segmentation methods as they do not use the audio as input during training. Second, current visual sound source localization baselines generate only coarse heatmaps for sounding objects instead of binary pixel-wise segmentation masks. To address the aforementioned challenges, our key idea is to apply multi-scale multiple-instance contrastive learning in audio-visual fusion to learn multi-scale cross-modal alignment, which differs from existing weakly-supervised semantic segmentation approaches. During training, we aim to use a reliable pseudo mask with multi-scale visual features as mask-level supervision for predicting accurate audio-visual segmentation masks in a weakly-supervised setting.

To this end, we propose a novel weakly-supervised audio-visual segmentation framework, namely WS-VAS, that can predict sounding source masks from both audio and image without using the ground-truth masks. Specifically, our WS-AVS leverages multi-scale multiple-instance contrastive learning in audio-visual fusion to capture multi-scale cross-modal alignment for addressing modality uncertainty in weakly-supervised semantic segmentation methods. Then, the pseudo mask refined by contrastive class-agnostic maps will serve as pixel-level guidance during training. Compared to previous visual sound source localization approaches, our method can generate binary accurate audio-visual segmentation masks.

Empirical experiments on AVSBench [1] comprehensively demonstrate the state-of-the-art performance against previous weakly-supervised baselines. In addition, qualitative visualizations of segmentation masks vividly showcase the effectiveness of our WS-AVS in predicting sounding object masks from both audio and image. Extensive ablation studies also validate the importance of audio-visual fusion with multi-scale multiple-instance contrastive learning and pseudo mask refinement by contrastive class-agnostic maps in single-source audio-visual segmentation.

Our main contributions can be summarized as follows:

- We investigate a new weakly-supervised multi-modal problem that predicts sounding object masks from both audio and image without needing pixel-level annotations.
- We present a novel framework for weakly-supervised audio-visual segmentation, namely WS-AVS, with multi-scale multiple-instance contrastive learning to capture multi-scale audio-visual alignment.
- Extensive experiments comprehensively demonstrate the state-of-the-art superiority of our WS-AVS over previous baselines on single-source sounding object segmentation.

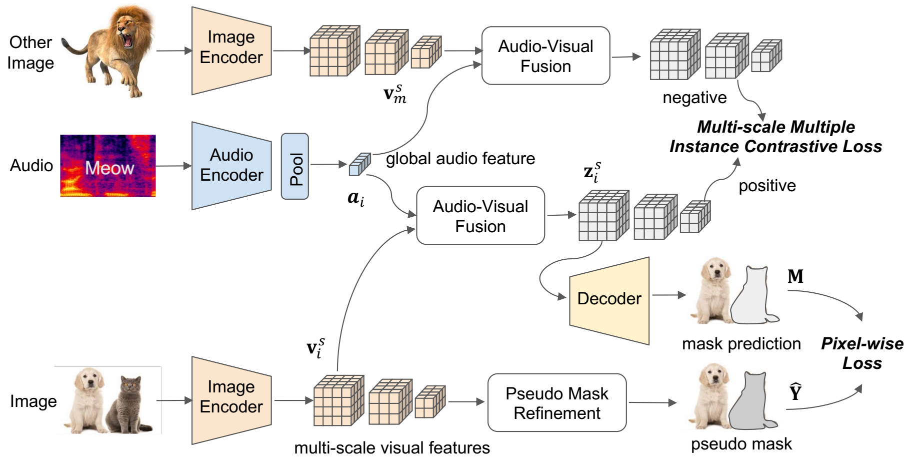

Figure 2: Illustration of the proposed Weakly-Supervised Audio-Visual Segmentation (WS-AVS) framework. Global audio features $\mathbf{a}_i$ and multi-scale visual features $\{\mathbf{v}_i^s\}_{s=1}^S$ are extracted from the audio spectrogram and image via an audio encoder and an image encoder, respectively. The Audio-Visual Fusion module combines multi-scale multiple-instance contrastive learning to capture multi-scale cross-modal alignment between global audio and multi-scale visual features. Then, with multi-scale visual features $\{\mathbf{v}_i^s\}_{s=1}^S$, the Pseudo Mask Refinement module generates a reliable pseudo mask $\hat{\mathbf{Y}}$ for sounding objects by using contrast class-agnostic map. Finally, a decoder is applied on multi-scale audio-visual features $\{\mathbf{z}_i^s\}_{s=1}^S$ to predict masks. The pixel-wise loss is simply optimized between the predicted mask $\mathbf{M}$ and the pseudo mask $\hat{\mathbf{Y}}$.

2 Related Work

Audio-Visual Learning. Audio-visual learning has been explored in many existing works [32, 33, 34, 35, 2, 36, 37, 38, 39, 40, 41] to learn audio-visual correspondence from those two distinct modalities in videos. Given video sequences, the goal is to push away those from different pairs while closing audio and visual representations from the same pair. Such cross-modal correspondence is useful for several tasks, such as audio/speech separation [38, 42, 43, 36, 37, 44, 45, 46, 47], visual sound source localization [2, 3, 4, 5, 6, 7, 8, 9, 10], audio spatialization [48, 49, 50, 39], and audio-visual parsing [51, 52, 53]. In this work, we mainly focus on learning audio-visual association for pixel-level audio-visual segmentation, which is more demanding than those tasks mentioned above.

Audio-Visual Segmentation. Audio-visual segmentation is a challenging problem that predicts pixel-wise masks for sounding objects in the image. In order to address this issue, Zhou *et al.* [1] introduced the first audio-visual segmentation benchmark with pixel-level annotations for each 5-second video, where a binary mask is used to indicate the pixels of sounding objects for the corresponding audio. Moreover, they proposed an encoder-decoder network with a temporal pixel-by-pixel audio-visual interaction module to solve the pixel-level audio-visual segmentation task. However, their approach requires exhausted pixel-level annotations as the main supervision. Without this supervision, their performance degrades significantly as observed in our experiments in Section 4.2. In this work, we aim to get rid of the dependency of pixel-level labels, by leveraging only instance-level labels for weakly-supervised audio-visual segmentation.

Weakly-Supervised Semantic Segmentation. Weakly-supervised visual segmentation aims at reducing the burden of collecting pixel-level annotations for image segmentation at a large scale required by its fully-supervised counterpart. Early methods explored different forms of weak supervision, including image-level label [54, 55], scribble [13, 15], and bounding box [12, 31, 56]. Typically, CAM [18] was introduced to generate class activation maps as pseudo pixel-level segmentation masks for weakly-supervised object localization. Based on CAM, following works tried to optimize the coarse class activation map for more precise object localization, by expanding localization regions [21, 22, 23], using stochastic inference [24], exploring boundary constraints [25, 26, 27], or mining and erasing regions of objects [28, 29]. More recently, C^{2}AM [30] leveraged contrastive learning with the semantic relation between foreground and background as positive and negative pairs to generate a class-agnostic activation map. Different from weakly-supervised visual segmentation baselines, we aim to develop a fully novel framework to aggregate semantics from both audio and visual representations with the only instance-level annotation. To the best of our knowledge, we are the first to explore class-agnostic activation maps on weakly-supervised audio-visual segmentation. In addition, we leverage audio-visual multi-scale fusion to boost the segmentation performance.

3 Method

Given a clip of audio and video frames, our target is to predict pixel-level masks for sounding objects in the frame. We present a novel Weakly-Supervised Audio-Visual Segmentation framework named WS-AVS for single source audio-visual segmentation without involving the pixel-wise annotation, which mainly consists of two modules, Audio-Visual Fusion with Multi-scale Multiple Instance Contrastive Learning in Section 3.2 and Pseudo Mask Refinement by contrastive class-agnostic maps in Section 3.3.

3.1 Preliminaries

In this section, we first describe the problem setup and notations, and then revisit the first work [1] with pixel-level annotations for single source audio-visual segmentation.

Problem Setup and Notations. Given an audio spectrogram and an image, our goal is to predict the binary segmentation mask for the sound source in the image spatially. Let $\mathcal{D} = (a_i, v_i) : i = 1, ..., N$ be a dataset of paired audio $a_i \in \mathbb{R}^{T\times F}$ and visual data $v_i \in \mathbb{R}^{3\times H\times W}$, where only one sound source a_i are assumed to be existing in v_i. Note that T, F denotes the dimension of time and frequency of the audio spectrogram. We follow previous work [1, 9] and first encode the audio and visual inputs using a two-stream neural network encoder and projection heads, denoted as $f_a(\cdot), g_a(\cdot)$ and $f_v(\cdot), g_v(\cdot)$ for the audio and images, separately. The audio encoder computes global audio representations $\mathbf{a}_i = g_a(f_a(a_i)), \mathbf{a}_i \in \mathbb{R}^{1\times D}$ and the visual encoder extracts multi-scale representations $\{\mathbf{v}_i^s\}_{s=1}^S = g_v(f_v(v_i)), \mathbf{v}_i^s \in \mathbb{R}^{D\times H^s\times W^s}$ for each sth stage. During the training, we do not have mask-level annotations $\mathbf{Y} \in \mathbb{R}^{H\times W}$.

Revisit Audio-Visual Segmentation with Pixel-level Annotations. To solve the single source audio-visual segmentation problem, AVS [1] introduced the pixel-wise audio-visual fusion module to encode the multi-scale visual features and global audio representations. After the cross-modal fusion, the audio-visual feature map $\mathbf{z}_i^s$ at the sth stage is updated as:

$$\mathbf{z}_i^s = \mathbf{v}_i^s + \mu\left(\frac{\theta(\mathbf{v}_i^s)\phi(\hat{\mathbf{a}}_i)^\top}{H^s W^s}\omega(\mathbf{v}_i^s)\right) \tag{1}$$

where $\hat{\mathbf{a}}_i \in \mathbb{R}^{D\times H^s\times W^s}$ denotes the duplicated version of the global audio representation $\mathbf{a}_i$ that repeats $H^s W^s$ times at the sth stage. $\mu(\cdot), \theta(\cdot), \phi(\cdot), \omega(\cdot)$ denote the $1\times 1\times 1$ convolution operator. Then those updated multi-stage feature maps are passed into the decoder of Panoptic-FPN [57] and a sigmoid activation layer to generate the final output mask $\mathbf{M} \in \mathbb{R}^{H\times W}$. With the pixel-level annotation $\mathbf{Y}$ as supervision, they applied the binary cross entropy loss between the prediction and label as:

$$\mathcal{L}_{\text{baseline}} = \text{BCE}(\mathbf{M}, \mathbf{Y}) \tag{2}$$

However, such a training mechanism is extremely dependent on pixel-level annotations. Without this supervision, their performance on single source audio-visual segmentation deteriorates significantly, as observed in Section 4.2. In the meanwhile, current weakly-supervised semantic segmentation (WSSS) [18, 30] and visual sound source localization (VSSL) [9] methods will pose two main challenges. First, there is no multi-modal constraint for pixel-level mask prediction in the previous WSSS methods as they do not involve the audio during training. Second, existing VSSL approaches only predict coarse heatmaps for sounding sources instead of accurate pixel-wise segmentation masks. To address these challenges, we propose a novel weakly-supervised audio-visual segmentation framework for single sounding object segmentation, which is composed of Audio-Visual Fusion with multi-scale multiple-instance contrastive learning and Pseudo Mask Refinement by contrastive class-agnostic map, as shown in Figure 2.

3.2 Audio-Visual Fusion with Multi-scale Multiple-Instance Contrastive Learning

In order to address the modality uncertainty brought by the previous WSSS baselines [18, 30], inspired by EZ-VSL [9], we propose to conduct audio-visual fusion by focusing only on the most aligned regions when matching the audio to the multi-scale visual feature. This is because most locations in the video frame do not correspond to the sounding object, and thus the representations at these locations should not be aligned with the audio during training.

Table 1: Comparison results (%) of weakly-supervised audio-visual segmentation. “ws” denotes the weakly-supervised baseline, where only the instance-level class is used for training.

Method	mIoU	F-score
AVS [1] (ws)	12.63	24.99
CAM [18]	19.26	27.88
EZ-VSL [9]	29.40	35.70
C^{2}AM [30]	30.87	36.55
WS-AVS (ours)	**34.13**	**51.76**

To align the audio and visual features at locations corresponding to sounding sources, we apply the multi-scale multiple-instance contrastive learning (M^{2}ICL) objective to align at least one location in the corresponding bag of multi-scale visual features with the audio representation in the same mini-batch, which is defined as:

$$\mathcal{L}_{a\rightarrow v} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{s=1}^{S}\log\frac{\exp\left(\frac{1}{\tau}\texttt{sim}(\mathbf{a}_i,\mathbf{v}_i^s)\right)}{\sum_{m=1}^{B}\exp\left(\frac{1}{\tau}\texttt{sim}(\mathbf{a}_i,\mathbf{v}_m^s)\right)} \tag{3}$$

where the similarity $\texttt{sim}(\mathbf{a}_i,\mathbf{v}^s)$ denotes the max-pooled audio-visual cosine similarity of $\mathbf{a}_i$ and $\mathbf{v}_i^s$ across all spatial locations at s stage. That is, $\texttt{sim}(\mathbf{a}_i,\mathbf{v}_i^s) = \max_{H^sW^s}(\mathbf{a}_i,\mathbf{v}_i^s)$. B is the batch size, D is the dimension size, and τ is a temperature hyper-parameter.

Similar to EZ-VSL [9], we use a symmetric loss of Eq. 4 to discriminate negative audio bags from other audio samples in the same mini-batch, which is defined as

$$\mathcal{L}_{v\rightarrow a} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{s=1}^{S}\log\frac{\exp\left(\frac{1}{\tau}\texttt{sim}(\mathbf{a}_i,\mathbf{v}_i^s)\right)}{\sum_{m=1}^{B}\exp\left(\frac{1}{\tau}\texttt{sim}(\mathbf{a}_m,\mathbf{v}_i^s)\right)} \tag{4}$$

where $\mathbf{a}_m$ denote the global audio visual from other sample m in the mini-batch. The overall audio-visual fusion objective with the multi-scale multiple-instance contrastive learning mechanism is defined as:

$$\mathcal{L}_{\text{avf}} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{s=1}^{S}\log\frac{\exp\left(\frac{1}{\tau}\texttt{sim}(\mathbf{a}_i,\mathbf{v}_i^s)\right)}{\sum_{m=1}^{B}\exp\left(\frac{1}{\tau}\texttt{sim}(\mathbf{a}_i,\mathbf{v}_m^s)\right)} + \log\frac{\exp\left(\frac{1}{\tau}\texttt{sim}(\mathbf{a}_i,\mathbf{v}_i^s)\right)}{\sum_{m=1}^{B}\exp\left(\frac{1}{\tau}\texttt{sim}(\mathbf{a}_m,\mathbf{v}_i^s)\right)} \tag{5}$$

Optimizing the loss will push the model to learn discriminatively global audio representation $\mathbf{a}_i$ and multi-scale visual features $\{\mathbf{v}_i^s\}_{s=1}^{S}$. Then, these features are fed forward into Eq. 1 to generate updated multi-scale audio-visual features $\{\mathbf{z}_i^s\}_{s=1}^{S}$. Finally, with these updated multi-scale audio-visual features, we apply the decoder of Panoptic-FPN [57] to generate the output mask $\mathbf{M}$. Note that, the recent audio-visual segmentation baseline [1] does not give any cross-modal constraint on the audio and visual representations in the audio-visual fusion, which causes a significant performance drop when removing the ground-truth masks during training.

3.3 Pseudo Mask Refinement by Contrastive Class-agnostic Maps

The second challenge of spatial uncertainty in VSSL approaches [9] requires us to utilize mask-level supervision for generating precise final output mask $\mathbf{M}$. To generate reliable pseudo masks of training sets, motivated by recent WSSL pipelines [30], we introduce the Pseudo Mask Refinement by applying a contrastive class-agnostic map on multi-scale visual features $\{\mathbf{v}_i^s\}_{s=1}^{S}$. Specifically, we utilize the contrastive loss in [30] to close the distance between the representations in positive pairs (foreground-foreground, background-background) and push away the representations in negative pairs (foreground-background). Then, we use the background activation maps as pseudo labels to further train a salient object detector [58] to predict the salient region $\mathbf{S} \in \mathbb{R}^{1\times H\times W}$ in the image. Finally, the predicted salient cues are concatenated with the initial class-agnostic map $\mathbf{A} \in \mathbb{R}^{D\times H\times W}$ to perform the argmax operator along the embedding dimension to generate the binary pseudo mask $\hat{\mathbf{Y}} \in \mathbb{R}^{H\times W}$, which is formulated as:

$$\hat{\mathbf{Y}} = \mathbf{L}\left[\arg\max_{1+D}([\mathbf{S};\mathbf{A}])\right] \tag{6}$$

where [;] denotes the concatenation operator. Note that since we are focusing on single-source segmentation, $\mathbf{L} = [\mathbf{L}_S;\mathbf{L}_A] \in \mathbb{R}^{(1+D)\times H\times W}$. $\mathbf{L}_S \in \mathbb{R}^{1\times H\times W}$ is the full label of zeros for the salient cues, while $\mathbf{L}_A \in \mathbb{R}^{D\times H\times W}$ is the full label of ones for the class-agnostic maps. With the

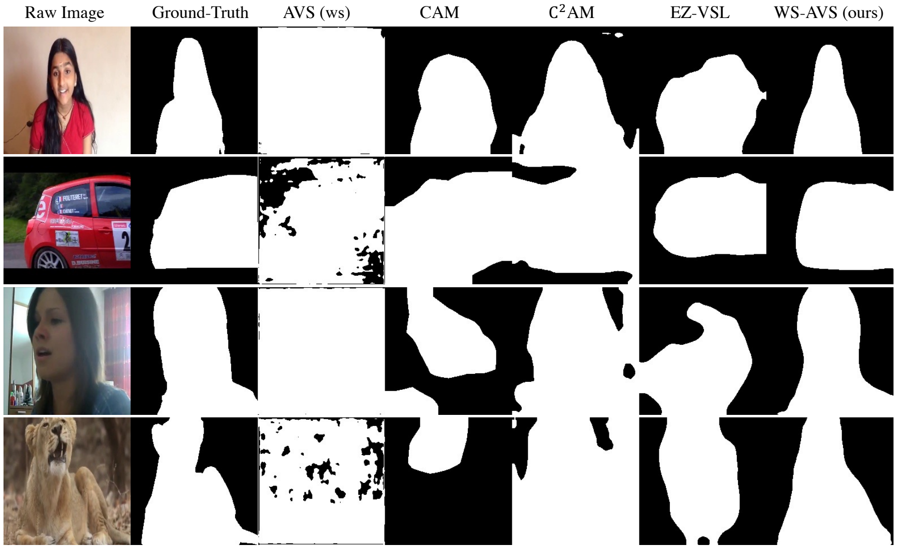

Figure 3: Qualitative comparisons with weakly-supervised audio-visual segmentation (AVS [1]), weakly-supervised semantic segmentation (CAM [18], C^{2}AM [30]), and visual sound source localization (EZ-VSL [9]) baselines on single-source segmentation. The proposed WS-AVS produces much more accurate and high-quality segmentation maps for sounding objects.

reliable pseudo mask refinement, we apply the binary cross entropy loss between the predicted mask $\mathbf{M}$ and the pseudo mask $\hat{\mathbf{Y}}$, which is defined as:

$$\mathcal{L}_{\text{pmr}} = \text{BCE}(\mathbf{M}, \hat{\mathbf{Y}}) \tag{7}$$

The overall objective of our model is simply optimized in an end-to-end manner as:

$$\mathcal{L} = \mathcal{L}_{\text{avf}} + \mathcal{L}_{\text{pmr}} \tag{8}$$

During inference, we follow previous work [1] and directly use the predicted mask $\mathbf{M} \in \mathbb{R}^{H \times W}$ as the final output. It is worth noting that the final localization map in the current VSSL methods [9] was generated through bilinear interpolation of the audio-visual feature map at the last stage.

4 Experiments

4.1 Experiment Setup

Datasets. AVSBench [1] contains 4,932 videos with 10,852 total frames from 23 categories including animals, humans, instruments, etc. Following prior work [1], we use the split of 3,452/740/740 videos for train/val/test in single source segmentation.

Evaluation Metrics. Following previous work [1], we use the averaged IoU (mIoU) and F-score to evaluate the audio-visual segmentation performance. mIoU computes the intersection-over-union (IoU) of the predicted mask and ground-truth mask for evaluating the region similarity. F-score calculates both the precision and recall for evaluating the contour accuracy.

Implementation. The input image is resized to a resolution of 224 × 224. The input audio takes the log spectrograms extracted from 3s of audio at a sample rate of 22050Hz. Following previous works [9, 10], we apply STFT to generate an input tensor of size 257 × 300 (257 frequency bands over 300 timesteps) using 50ms windows with a hop size of 25ms. We follow the prior audio-visual segmentation work [1] and use the ResNet50 [59] as the audio and visual encoder. The visual model is initialized using weights pre-trained on ImageNet [60]. The model is trained with the Adam optimizer with default hyper-parameters β_1 = 0.9, β_2 = 0.999, and a learning rate of 1e-4. The model is trained for 20 epochs with a batch size of 64.

4.2 Comparison to Prior Work

In this work, we propose a novel and effective framework for weakly-supervised audio-visual segmentation. To validate the effectiveness of the proposed WS-AVS, we comprehensively compare it to previous audio-visual segmentation, weakly-supervised semantic segmentation, and visual sound source localization baselines: 1) AVS [1](ws): the weakly-supervised version of the first audio-visual

Table 2: Ablation studies on Audio-Visual Fusion (AVF) and Pseudo Mask Refinement (PMR).

AVF	PMR	mIoU (↑)	F-score (↑)
✗	✗	12.63	24.99
✓	✗	31.85 (**+19.22**)	39.78 (**+14.79**)
✗	✓	32.46 (**+19.83**)	43.92 (**+18.93**)
✓	✓	**34.13** (**+21.50**)	**51.76** (**+26.77**)

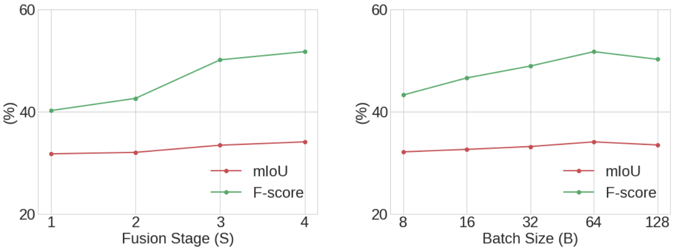

Figure 4: Effect of fusion stages and batch size on mIoU and F-score.

segmentation work, where we removed the pixel-level annotations for training; 2) CAM [18]: the first baseline using class-activation maps as pseudo segmentation masks for weakly-supervised object localization; 3) EZ-VSL [9]: the state-of-the-art visual sound source localization approach with coarse source maps as output; 4) C^{2}AM [30]: the very recent work for weakly-supervised semantic segmentation with only the image as input.

The quantitative comparison results are reported in Table 1. As can be seen, we achieve the best performance in terms of all metrics compared to previous weakly-supervised baselines. In particular, the proposed WS-AVS significantly outperforms AVS [1](ws), the weakly-supervised version of the state-of-the-art audio-visual segmentation approach, by 21.50 mIoU and 26.77 F-score. Moreover, we achieve superior performance gains of 3.26 mIoU and 15.21 F-score compared to C^{2}AM [30], which implies the importance of the proposed multi-scale multiple-instance contrastive learning for addressing the modality uncertainty in audio-visual fusion. Meanwhile, our WS-AVS outperforms EZ-VSL [9], the current state-of-the-art visual sound source localization baseline, by a large margin, where we achieve the performance gains of 4.73 mIoU and 16.06 F-score. These significant improvements demonstrate the superiority of our method in single-source audio-visual segmentation.

In order to qualitatively evaluate the localization maps, we compare the proposed WS-AVS with weakly-supervised AVS [1], CAM [18], C^{2}AM [30], and EZ-VSL [9] on single-source segmentation in Figure 3. From comparisons, three main observations can be derived: 1) Without pixel-level annotations, AVS [1], the first audio-visual segmentation fails to predict the masks for sounding objects; 2) the quality of segmentation masks generated by our method is much better than the strong visual sound source baseline, EZ-VSL [9]; 3) the proposed WS-AVS achieves competitive even better results on predicted masks against the weakly-supervised semantic segmentation baseline [30] by using contrast class-agnostic maps for prediction. These visualizations further showcase the superiority of our simple WS-AVS with multi-scale multiple-instance contrastive learning to guide segmentation for predicting accurate source masks.

4.3 Experimental Analysis

In this section, we performed ablation studies to demonstrate the benefit of introducing the Audio-Visual Fusion with multi-scale multiple-instance contrastive learning and Pseudo Mask Refinement module by contrastive class-agnostic map. We also conducted extensive experiments to explore the effect of fusion stages and batch size on weakly-supervised audio-visual segmentation. Furthermore, we visualized high-quality pseudo masks from Pseudo Mask Refinement in weakly-supervised training.

Audio-Visual Fusion & Pseudo Mask Refinement. To validate the effectiveness of the introduced audio-visual fusion with multi-scale multiple-instance contrastive learning (AVF) and Pseudo Mask Refinement by contrastive class-agnostic map (PMR), we ablate the necessity of each module and report the quantitative results in Table 2. We can observe that adding AVF to the vanilla baseline highly increases the results of single-source audio-visual segmentation by 19.22 mIoU and 14.79 F-score, which demonstrates the benefit of multi-scale multiple-instance contrastive learning (M^{2}ICL)

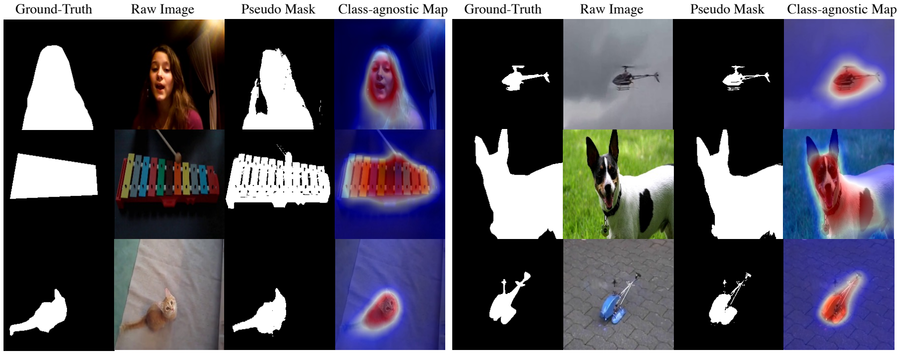

Figure 5: Qualitative comparisons of the ground-truth mask, pseudo mask, and class-agnostic map. Note that the generated pseudo mask is used for weakly-supervised training.

in extracting aligned cross-modal feature in fusion for source segmentation. Meanwhile, introducing only PMR in the baseline also increases the segmentation performance in terms of all metrics. More importantly, incorporating AVF with M^{2}ICL and PMR together into the baseline significantly raises the performance by 21.50 mIoU and 26.77 F-score. These improving results validate the importance of AVF with M^{2}ICL and PMR with contrastive class-agnostic map in addressing both modality and spatial challenges for generating accurate audio-visua masks.

Effect of Fusion Stages and Batch Size. The number of fusion stages and batch size used in the proposed M^{2}ICL affect the extracted cross-modal representations for audio-visual segmentation. To explore such effects more comprehensively, we varied the number of fusion stages from $\{1, 2, 3, 4\}$ and ablated the batch size from $\{8, 16, 32, 64, 128\}$. The comparison results of segmentation performance are shown in Figure 4. When the number of fusion stages is 4 and using batch size of 64 in M^{2}ICL, we achieve the best segmentation performance in terms of all metrics. With the increase of fusion stages from 1 to 4, the proposed WS-AVS consistently raises results, which shows the importance of multi-scale visual features in audio-visual fusion for learning discriminative cross-modal representations. Regarding the batch size, the performance of the proposed WS-AVS climbs with the increase of the batch size from 8 to 64. However, increasing the batch size from 64 to 128 will not continually improve the result since there might be some false negatives in the mini-batch for this training set with a relatively smaller size.

Generated Pseudo Mask. Generating reliable pseudo masks with contrast class-agnostic maps is critical for us to train the weakly-supervised framework. To better evaluate the quality of generated pseudo masks, we visualize the pseudo mask and class-agnostic map in Figure 5. As can be observed in the last column, the class-agnostic map can successfully localize the sounding object in the image. Furthermore, the pseudo masks in the third column predicted by Eq. 6 have high quality and they are even fine-grained compared to the manually annotated ground-truth masks. These meaningful visualization results further showcase the success of the Pseudo Mask Refinement in extracting reliable pseudo masks as guidance for single-source audio-visual segmentation.

4.4 Limitation

Although the proposed WS-AVS achieves superior results on single-source audio-visual segmentation, the performance gains of our approach on the mIoU metric are not significant. One possible reason is that our model easily overfits across the training phase, and the solution is to incorporate dropout and momentum encoders together for weakly-supervised audio-visual segmentation. Meanwhile, we notice that our model performs worse on mixed sound sources, such as a scenario with both female singing and playing tabla. The future work could be to add separation objectives or assign semantic labels to both audio and visual segments in audible videos.

5 Conclusion

In this work, we successfully investigate the weakly-supervised audio-visual segmentation to eliminate pixel-level annotations as supervision. Then we present WS-AVS, a novel Weakly-Supervised Audio-Visual Segmentation framework with multi-scale multiple-instance contrastive learning for capturing multi-scale audio-visual alignment. Furthermore, we conduct extensive experiments on AVSBench to demonstrate the superiority of our WS-AVS against previous weakly-supervised audio-visual segmentation, weakly-supervised semantic segmentation, and visual sound source localization approaches.

References

[1] 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 *European Conference on Computer Vision*, 2022. 1, 2, 3, 4, 5, 6, 7

[2] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4358–4366, 2018. 1, 2, 3

[3] Andrew Rouditchenko, Hang Zhao, Chuang Gan, Josh H. McDermott, and Antonio Torralba. Self-supervised audio-visual co-segmentation. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2357–2361, 2019. 1, 3

[4] Di Hu, Feiping Nie, and Xuelong Li. Deep multimodal clustering for unsupervised audiovisual learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9248–9257, 2019. 1, 3

[5] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. Self-supervised learning of audio-visual objects from video. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 208–224, 2020. 1, 3

[6] Janani Ramaswamy and Sukhendu Das. See the sound, hear the pixels. *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2959–2968, 2020. 1, 3

[7] Rui Qian, Di Hu, Heinrich Dinkel, Mengyue Wu, Ning Xu, and Weiyao Lin. Multiple sound sources localization from coarse to fine. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 292–308, 2020. 1, 3

[8] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Localizing visual sounds the hard way. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16867–16876, 2021. 1, 3

[9] Shentong Mo and Pedro Morgado. Localizing visual sounds the easy way. *arXiv preprint arXiv:2203.09324*, 2022. 1, 2, 3, 4, 5, 6, 7

[10] Shentong Mo and Pedro Morgado. A closer look at weakly-supervised audio-visual source localization. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 1, 3, 6

[11] George Papandreou, Liang-Chieh Chen, Kevin P Murphy, and Alan L Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1742–1750, 2015. 1

[12] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1635–1643, 2015. 1, 3

[13] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3159–3167, 2016. 1, 3

[14] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. What's the point: Semantic segmentation with point supervision. In *European conference on computer vision*, pages 549–565. Springer, 2016. 1

[15] Paul Vernaza and Manmohan Chandraker. Learning random-walk label propagation for weakly-supervised semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7158–7166, 2017. 1, 3

[16] Qibin Hou, Daniela Massiceti, Puneet Kumar Dokania, Yunchao Wei, Ming-Ming Cheng, and Philip HS Torr. Bottom-up top-down cues for weakly-supervised semantic segmentation. In *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 263–277. Springer, 2017. 1

[17] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1568–1576, 2017. 1

[18] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 1, 2, 3, 4, 5, 6, 7

[19] Tianfei Zhou, Wenguan Wang, Si Liu, Yi Yang, and Luc Van Gool. Differentiable multi-granularity human representation learning for instance-aware human semantic parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1622–1631, 2021. 1

[20] Jiapei Feng, Xinggang Wang, and Wenyu Liu. Deep graph cut network for weakly-supervised semantic segmentation. *Science China Information Sciences*, 64(3):1–12, 2021. 1

[21] Alexander Kolesnikov and Christoph H Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *European conference on computer vision*, pages 695–711. Springer, 2016. 1, 3

[22] Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu, and Jingdong Wang. Weakly-supervised semantic segmentation network with deep seeded region growing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7014–7023, 2018. 1, 3

[23] Yunchao Wei, Huaxin Xiao, Honghui Shi, Zequn Jie, Jiashi Feng, and Thomas S Huang. Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7268–7277, 2018. 1, 3

[24] Jungbeom Lee, Eunji Kim, Sungmin Lee, Jangho Lee, and Sungroh Yoon. Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5267–5276, 2019. 1, 3

[25] Wataru Shimoda and Keiji Yanai. Self-supervised difference detection for weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5208–5217, 2019. 1, 3

[26] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12275–12284, 2020. 1, 3

[27] Yu-Ting Chang, Qiaosong Wang, Wei-Chih Hung, Robinson Piramuthu, Yi-Hsuan Tsai, and Ming-Hsuan Yang. Weakly-supervised semantic segmentation via sub-category exploration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8991–9000, 2020. 1, 3

[28] Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. Tell me where to look: Guided attention inference network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9215–9223, 2018. 1, 3

[29] Qibin Hou, PengTao Jiang, Yunchao Wei, and Ming-Ming Cheng. Self-erasing network for integral object attention. *Advances in Neural Information Processing Systems*, 31, 2018. 1, 3

[30] Jinheng Xie, Jianfeng Xiang, Junliang Chen, Xianxu Hou, Xiaodong Zhao, and Linlin Shen. C2am: Contrastive learning of class-agnostic activation map for weakly supervised object localization and semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 989–998, 2022. 2, 3, 4, 5, 6, 7

[31] Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 876–885, 2017. 2, 3

[32] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2016. 3

[33] Andrew Owens, Jiajun Wu, Josh H. McDermott, William T. Freeman, and Antonio Torralba. Ambient sound provides supervision for visual learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 801–816, 2016. 3

[34] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 609–617, 2017. 3

[35] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 3

[36] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 570–586, 2018. 3

[37] Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba. The sound of motions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1735–1744, 2019. 3

[38] Chuang Gan, Deng Huang, Hang Zhao, Joshua B. Tenenbaum, and Antonio Torralba. Music gesture for visual sound separation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10478–10487, 2020. 3

[39] Pedro Morgado, Yi Li, and Nuno Nvasconcelos. Learning representations from audio-visual spatial alignment. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, pages 4733–4744, 2020. 3

[40] Pedro Morgado, Ishan Misra, and Nuno Vasconcelos. Robust audio-visual instance discrimination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12934–12945, 2021. 3

[41] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12475–12486, June 2021. 3

[42] Ruohan Gao, Rogerio Feris, and Kristen Grauman. Learning to separate object sounds by watching unlabeled video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 35–53, 2018. 3

[43] Ruohan Gao and Kristen Grauman. Co-separating sounds of visual objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3879–3888, 2019. 3

[44] Ruohan Gao, Tae-Hyun Oh, Kristen Grauman, and Lorenzo Torresani. Listen to look: Action recognition by previewing audio. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10457–10467, 2020. 3

[45] Yapeng Tian, Di Hu, and Chenliang Xu. Cyclic co-learning of sounding object visual grounding and sound separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2745–2754, 2021. 3

[46] Ruohan Gao and Kristen Grauman. Visualvoice: Audio-visual speech separation with cross-modal consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15495–15505, 2021. 3

[47] Tanzila Rahman and Leonid Sigal. Weakly-supervised audio-visual sound source detection and separation. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021. 3

[48] Pedro Morgado, Nuno Nvasconcelos, Timothy Langlois, and Oliver Wang. Self-supervised generation of spatial audio for 360°video. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 3

[49] Ruohan Gao and Kristen Grauman. 2.5d visual sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 324–333, 2019. 3

[50] Changan Chen, Unnat Jain, Carl Schissler, S. V. A. Garí, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. Soundspaces: Audio-visual navigation in 3d environments. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 17–36, 2020. 3

[51] Yapeng Tian, Dingzeyu Li, and Chenliang Xu. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In *Proceedings of European Conference on Computer Vision (ECCV)*, page 436–454, 2020. 3

[52] Yu Wu and Yi Yang. Exploring heterogeneous clues for weakly-supervised audio-visual video parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1326–1335, 2021. 3

[53] 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. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 3

[54] Binglu Wang, Yongqiang Zhao, and Xuelong Li. Multiple instance graph learning for weakly supervised remote sensing object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–12, 2021. 3

[55] Dong Zhang, Hanwang Zhang, Jinhui Tang, Xian-Sheng Hua, and Qianru Sun. Causal intervention for weakly-supervised semantic segmentation. *Advances in Neural Information Processing Systems*, 33:655–666, 2020. 3

[56] Youngmin Oh, Beomjun Kim, and Bumsub Ham. Background-aware pooling and noise-aware loss for weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6913–6922, 2021. 3

[57] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollar. Panoptic feature pyramid networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6399–6408, 2019. 4, 5

[58] Jiang-Jiang Liu, Qibin Hou, Ming-Ming Cheng, Jiashi Feng, and Jianmin Jiang. A simple pooling-based design for real-time salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3917–3926, 2019. 5

[59] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 6

[60] Jia Deng, Wei Dong, Richard Socher, Li-Jia. Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009. 6

Appendix

In this appendix, we provide the detailed resource, work division, and milestones for this work.

Resources, Work Division, and Milestone

We have two NVIDIA Tesla V100 GPUs (32GB) and one NVIDIA RTX3090 (24GB) for experiments.

Shentong Mo: edit reports, review literature, and implement the proposed WS-AVS.

Yixuan Chen: edit reports, review literature, and run weakly-supervised visual segmentation baseline.

Ding Bai: edit reports, review literature, and run audio-visual segmentation baseline.

Week 9	●	Submit the project proposal and reproduce fully-supervised baseline. ✓
Week 11	●	Read more related works and run weakly-supervised visual segmentation baseline. ✓
Week 13	●	Implement our proposed WS-AVS and run experiments. ✓
Week 15	●	Wrap up experimental results and finish the final report. ✓