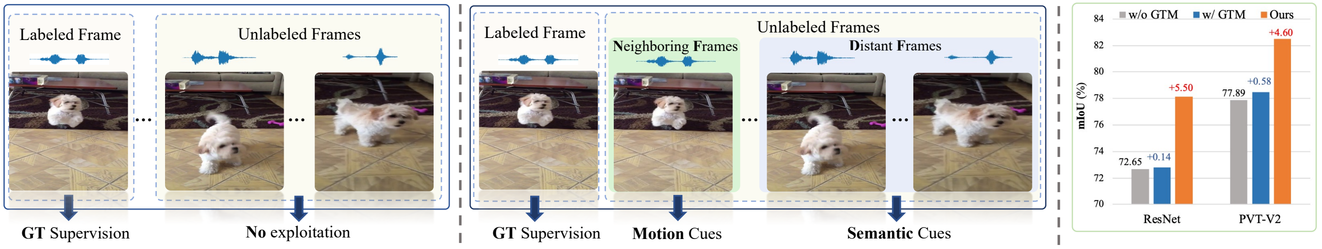
Audio-Visual Segmentation via Unlabeled Frame Exploitation

Jinxiang Liu[1], Yikun Liu[1], Fei Zhang[1], Chen Ju[1], Ya Zhang[1,2]✉, Yanfeng Wang[1,2]
[1] Cooperative Medianet Innovation Center, Shanghai Jiao Tong University [2] Shanghai AI Laboratory
{jinxliu, yikunliu, ferenas, ju_chen, ya_zhang, wangyanfeng622}@sjtu.edu.cn



(a) Previous methods (w/ GTM)

(b) Our proposed method (Ours)

(c) Performance comparison

Figure 1. **Comparison between previous methods and ours on how to harness the unlabeled frames.** (a) Previous methods perform global temporal modeling (GTM) to process all frames from a sequence including labeled and unlabeled ones, without the exploitation of the unlabeled frames. (b) Our method employs two types of unlabeled frames: (i) the *neighboring frames* (**NFs**) provide motion cues for accurately segmenting the sounding object; (ii) the *distant frames* (**DFs**) contain semantic cues for enhancing data diversity. (c) Based on TPAVI method, compared to the model trained only using labeled frames (w/o GTM), previous methods using global temporal modeling (w/ GTM) only show marginal performance gain; while our method achieves significant improvement with the unlabeled frames.

Abstract

Audio-visual segmentation (AVS) aims to segment the sounding objects in video frames. Although great progress has been witnessed, we experimentally reveal that current methods reach marginal performance gain within the use of the unlabeled frames, leading to the underutilization issue. To fully explore the potential of the unlabeled frames for AVS, we explicitly divide them into two categories based on their temporal characteristics, i.e., neighboring frame (NF) and distant frame (DF). NFs, temporally adjacent to the labeled frame, often contain rich motion information that assists in the accurate localization of sounding objects. Contrary to NFs, DFs have long temporal distances from the labeled frame, which share semantic-similar objects with appearance variations. Considering their unique characteristics, we propose a versatile framework that effectively leverages them to tackle AVS. Specifically, for NFs, we exploit the motion cues as the dynamic guidance to improve the objectness localization. Besides, we exploit the semantic cues in DFs by treating them as valid augmentations to the labeled frames, which are then used to enrich data diversity in a self-training manner. Extensive experimental results demonstrate the versatility and superiority of our method, unleashing the power of the abundant unlabeled frames.

1. Introduction

Humans perceive the surroundings not only by seeing but also by hearing to accurately and efficiently obtain the target information [23]. In the audio-visual understanding field, the demand to visually attend to the auditory objects has driven the exploration of the audio-visual segmentation (AVS) task [81]. The goal of AVS is to localize and segment the sounding objects in the video frames with the guidance of audio signals. And the successful grounding of auditory objects with the AVS task will benefit a wide range of downstream tasks such as multi-modal content editing [1, 42, 43], video surveillance [6, 7], and robot industry [17, 74].

To address the task, current methods [19, 44, 47, 49, 50, 54, 81] are based on the dataset which is *sparsely* annotated. Concretely, due to the high labeling costs, only few frames in a video frame sequence are annotated with groundtruth masks, leaving the rest abundance of frames unlabeled. For example, in AVSBench-S4 dataset [81], only one sampled frame is annotated for a 5-second video. Despite the predominance of unlabeled frames within the datasets, current approaches [19, 44, 47, 49, 54, 81] adopt global temporal modeling module (GTM) that overemphasizes on exploiting the labeled frames to help address AVS, which may lead to the *underutilization* of the abundant unlabeled frames. To further verify this, we perform experiments based on the typical TPAVI [81] method. The results in Fig. 1 (c) demonstrate that compared with the baseline model (w/o GTM) trained with only labeled frames, current approach (w/ GTM) without tailored handling for the unlabeled frames only provides marginal improvement. Therefore, we are motivated to explore a more effective way to utilize the unlabeled frames for the AVS task.

Before delving into the exploitation of the unlabeled frames, let us rethink the characteristics of the abundant unlabeled frames. Taking Fig. 1 (b) as an example, given a target labeled frame describing "dog jumping", its neighboring unlabeled frames usually have very tiny visual appearance changes. For the distant frames, they usually contain the same object but with large appearance variations to the object in the labeled frame, *e.g.*, the dog has transformed from the pose "jumping" in the labeled frame to "walking" or "standing still" in the distant frames. Based on the observation, we start by first dividing the unlabeled frames into two categories: *neighboring frame* (NF) and *distant frame* (DF), based on the temporal distance with the target labeled frame. Though the visual changes are very limited, NFs often contain rich dynamic motion information that is important to the audio-visual understanding [2, 8, 12, 64, 78]. If properly used, the motion can not only assist in the accurate localization of the sounding objects but also provide the shape details of objects. For the DFs, both they and the labeled frame reflect the different stages of an audio-visual event [22, 23, 53, 62]. Contrary to the NFs, this long-term temporal relationship means that the DFs generally share the same or semantic-similar objects but with large appearance variations. Therefore, DFs could serve as the natural *semantic augmentations* for the labeled frames, which can be utilized to diversify the training data, thereby enhancing the model generalization capabilities.

Considering the characteristics of NFs and DFs, we propose a universal *unlabeled frame exploitation (UFE)* framework to leverage the two types of unlabeled frames with different strategies. For NFs, we extract the motion by calculating the optical flow between the target labeled frame and its NFs. And we explicitly feed the flow as model input to incorporate the motion guidance, which is complementary to the still RGB frame. In terms of DFs, since they are the natural *semantic augmentations* to labeled frames, the training data could be significantly enriched beyond the labeled frames. To this end, we propose a teacher-student network training framework to provide valid supervision for the unlabeled frames with the *weak-to-strong consistency*, where the predictions for the strong-augmented frames from the student are supervised by the predictions for the weak-augmented ones from the teacher. We perform the experiments by applying our proposed framework to two representative methods TPAVI [81] and AVSegFormer [19]. Extensive experimental results demonstrate the effectiveness of our proposed method to attack the AVS task by exploiting the unlabeled frames. The main contributions are:

- We propose a simple but effective partition strategy for the unlabeled frames based on the temporal characteristics, i.e., *neighboring frames* and *distant frames*, relieving the *underutilization* issue in AVS.
- We propose UFE, a versatile framework that leverages the NFs and DFs, where NFs provide motion guidance and DFs enhance the data diversity beyond the labeled frames, explicitly improving the objectness segmentation.
- Extensive experiments show our method can effectively exploit the abundant unlabeled frames and achieves new state-of-the-art performance on the AVS task, *e.g.*, 78.96 mIoU with ResNet backbone and 83.15 mIoU with PVT backbone on AVSBench-S4 dataset.

2. Related Work

Audio-Visual Segmentation. With the advancement of multi-modal learning [4, 33, 35–37], many audio-visual understanding problems have been studied, such as audio-visual sound separation [9, 18, 70, 77, 79], audio-visual segmentation [3, 27–29, 45, 48, 50, 60, 61, 66] and audio-visual video understanding [39, 41, 46, 69, 69]. In this paper, we focus on the audio-visual segmentation (AVS) task, whose purpose is to segment the sounding objects in video frames. Previous methods [3, 27–29, 45, 48, 59–61, 66] usually tackled the task in self-supervised or weakly-supervised learning and termed the task as visual sound source localization. Recently researchers [19, 44, 47, 49, 50, 54, 81] have been tackling AVS under the umbrella of supervised learning on the *sparsely-annotated* AVSBench dataset [81]. And based on architecture, we divide these methods into two categories: FCN-based [54, 81] and transformer-based methods [19, 44, 47, 49, 50]. For the FCN-based methods, the typical model is TPAVI [81]. The key design of [81] is the temporal pixel-wise audio-visual interaction (TPAVI) module which performs audio-visual feature fusion similar to the non-local block [73]. In terms of the transformer-based models [19, 44, 47, 49, 50], the general idea is to achieve audio-visual fusion and mask decoding with transformer [71]. For example, AVSegFormer [19] employs cross-attention to merge the spatial-temporal mask visual features and the audio embeddings. Although the architecture designs of FCN-based and transformer-based methods differ, they share a similar architecture pipeline, which consists of feature extraction, audio-visual fusion, and mask decoding. Besides, we observe current methods usually process the labeled frames and unlabeled frames from a video equally and predict the segmentations for all frames. However, due to the lack of annotations for the unlabeled data, the predictions for the unlabeled frames have *no* supervision. Inevitably, the methods without special handling for unlabeled frames lead to a suboptimal utilization problem.

Motion and Sound. Object motions and air vibration cause sound. We humans usually perceive sound together with the motion of visual objects. This strong relation between sound and motion has been studied by previous methods, *e.g.*, [14, 40] employed probabilistic models to investigate the relationship between motion and sound. Moreover, many previous works [2, 5, 8, 11, 12, 16, 58, 64, 78] have shown the important role that motion plays in audio-visual learning. For example, lip motion is a vital clue for speech processing tasks such as speech denoising [16] and speech separation [8, 58]. Other studies [2, 78] also modelled the temporal motion information [31, 32, 34, 80] in visual frames to solve the cocktail-party problem. However, few works explore motion for the AVS task.

Teacher-Student Network. Teacher-student network has become the dominant architecture for many problems [15, 21, 26, 30, 38, 51, 65, 68, 75, 76, 82]. In a teacher-student network, the prediction of the teacher model is used to regularize the prediction of the student model, thereby, transferring the teacher's knowledge. When employed to handle scenarios encompassing both labeled and unlabeled data, the teacher trained with labeled data can generate pseudo-labels for the unlabeled data, which the student then tries to match. In our model design, we first separate the labeled frames and unlabeled frames from a sequence for independent processing, and frame-wisely consider the labeled and the unlabeled frames across the dataset in our framework. To exploit the unlabeled frames, we adopt the weak-to-strong consistency [51, 55, 65, 76] by regularizing the predictions for the strong-augmented unlabeled frames from the student with the predictions for the weak-augmented unlabeled frames from the teacher. Thereby, the unlabeled frames can obtain supervision in a self-supervised manner.

3. Motivation

For the task of audio-visual segmentation (AVS), the input data consists of a sequence of sampled video frames $\mathcal{V} = \{I_i\}_{i=1}^{T}$, where $I_i \in \mathbb{R}^{3\times H_0\times W_0}$, and its corresponding audios $\mathcal{A} = \{a_i\}_{i=1}^{T}$, where $a_i \in \mathbb{R}^d$ is the audio clip with each video frame. The objective of the AVS task is to segment the sounding objects corresponding to the its audio in each frame in $\mathcal{V}$. The target segmentation is the binary masks for each frame $\mathcal{Y} = \{y_i\}_{i=1}^{T}$, where $y_i \in \{0,1\}^{H_0\times W_0}$.

As mentioned in Sec. 2, current mainstream approaches to AVS falls into two categories: the FCN-based methods [54, 81] and transformer-based methods [19, 44, 47, 49, 50]. Though the methods vary, the pipelines of methods can be abstracted into three subsequent steps: feature extraction with image encoder Φ_{image} and audio encoder Φ_{audio}, multi-modal feature fusion with fusion module Φ_{fusion}, and mask prediction with mask decoder Φ_{dec}. Formally, the predicted segmentation $\mathcal{P}$ is obtained with:

$$\mathcal{P} = \Phi_{\text{dec}}\left(\Phi_{\text{fuse}}\left(\Phi_{\text{image}}\left(\mathcal{V}\right), \Phi_{\text{audio}}\left(\mathcal{A}\right)\right)\right). \quad (1)$$

Notably, mainstream methods treat the labeled frames and unlabeled frames sampled from a video sequence equally and predict the masks for all frames. However, only the labeled frames have groundtruth supervision while the remaining abundant unlabeled frames have *no* supervision. And the only possible benefit which the unlabeled frames might provide for labeled frames is the contextual information with the global temporal modeling (GTM) operation. Concretely, global temporal modeling (GTM) employs cross-attention [71] to model the temporal relationships of the features across all the frames from a video, including labeled and unlabeled ones. To illustrate, Zhou et al. [81] deployed the cross-attention to integrate the space-time relations of the features in the TPAVI module in the audio-visual fusion stage; likewise, Gao et al. [19] proposed the channel-attention mixer based on the cross-attention in the audio-visual fusion stage to obtain the mask features.

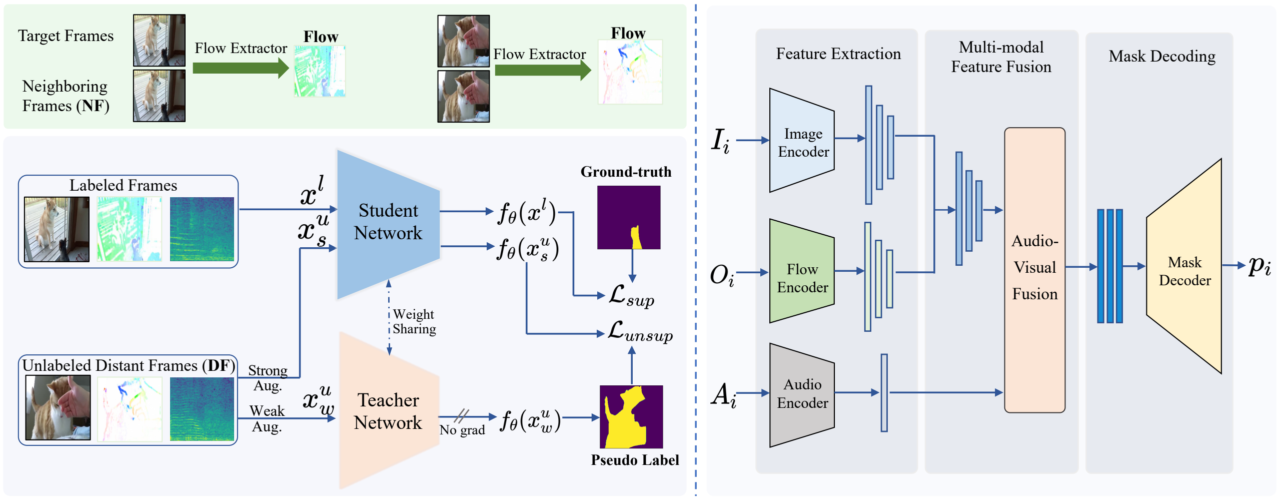
To measure the improvement by exploiting the unlabeled frames with GTM of the previous method [19, 81], we establish the baseline by discarding the unlabeled frames and only using the labeled frames for model training. We perform experiments on AVSBench-S4 dataset and compare the performance with two typical methods TPAVI [81] and AVSegFormer [19]. The results on TPAVI baseline model are shown in Fig. 1 (c), compared to the model trained with only labeled frames (w/o GTM), previous method [81] based on global temporal modeling (w/ GTM) achieves only marginal performance gain: 0.14 gain with ResNet and 0.58 gain with PVT in mIoU ($\mathcal{M}_{\mathcal{J}}$). For more metrics and more results on the AVSegFormer baseline method, please refer to the supplementary materials.

The results demonstrate the major issue of current methods: the *underutilization* of the unlabeled frames to boost the performance for the AVS task. Based on the observation, we intend to devise a more effective method to fully exploit the unlabeled frames, which is elaborated as follows.

4. Method

4.1. Framework Overview

Technically, our proposed framework can be established based on either FCN-based methods or transformer-based methods. Since both categories of methods are divided into three steps as elaborated in Sec. 2, our proposed framework also has three steps in the inference stage as illustrated in Fig. 2 (b). The model accepts the image I_i, the calculated flow O_i and its audio A_i as model inputs, then goes through three successive steps, to predict the target segmentation mask. The main difference from previous methods is that we harness the optical flow O_i extracted from the target frame and its unlabeled *neighboring frame* (NF) as model



(a) Teacher-Student Network Training

(b) Inference Pipeline

Figure 2. Overview of our framework to exploit unlabeled frames. (a) **Teacher-student network for training**. Student network is optimized with $\mathcal{L}_{sup}$ and $\mathcal{L}_{unsup}$. $\mathcal{L}_{sup}$ is computed with the predicted mask $f_\theta(x^l)$ and its groundtruth for the labeled frame x^l; $\mathcal{L}_{unsup}$ is computed between $f_\theta(x_s^u)$ from the student and the predicted pseudo mask $f_\theta(x_w^u)$ for the strong-augmented unlabeled image from teacher. (b) **Inference pipeline** of the framework. We incorporate flow as auxiliary input to exploit the motion cues within NFs.

input, in order to exploit the motion information to guide the model to focus on the exact sounding object thereby achieving accurate segmentation.

Regarding the exploitation of *distant frames* (DFs), we consider leveraging the semantic information of the distant frames and integrating the abundance of DFs to enrich the training data diversity. Even though DFs have *no* groundtruth annotations, we adopt the teacher-student network in the training phase, as shown in Fig. 2 (a). Different from previous methods where the unlabeled frames receive *no* supervision, our teacher-student network can provide supervision for the unlabeled distant frames to exploit the unlabeled frames with the *weak-to-strong consistency* from pseudo-labeling.

4.2. Neighboring Frame Exploitation

Sound occurs due to object motions and air vibration, thus sound has strong associations with motions. For example, when someone speaks, it always comes along with the movement of the speaker's lips; the sound of a musical instrument usually comes along with the hand movement of the player. The motion information can provide important dynamic cues for achieving the AVS task, which is complementary to the static information that still RGB frame provides. Concretely, i). motion can assist in localizing the exact sounding object and resolve the identity ambiguity. ii). motion information can even outline the shape and contour of the sounding objects, which contributes to accurate segmentations with better fine-grained details. And the motion information of the target frame can be simply extracted by exploiting the neighboring unlabeled frames. Concretely, we use the the target frame and its temporally-adjacent unlabeled frame in raw frame sequence to compute the optical flow, which serves a common way for motion estimation [13, 63]. And we leverage the optical flow by incorporating it as model input together with the target frame.

Specifically, for the ith sampled target frame from a video to be sent into the model, we utilize the RGB frame I_i^l and its subsequent neighboring unlabeled frame I_i^n to calculate the optical flow O_i with Gunnar Farneback algorithm [10]. Then as shown in Fig. 2 (b) we extract image and optical flow features with the image encoder and optical flow encoder separately. Following [19, 81], the image encoder can be either ResNet50 [24] or PVT-v2 [72]. We represent the extracted multi-scale image features as $\mathcal{F}_{vi} \in \mathbb{R}^{h_i \times w_i \times C_i}$, where $(h_i, w_i) = (H, W)/2^{i+1}$ for $i = 1, ..., 4$. For the optical flow encoder, we adapt the pretrained ResNet-18 architecture by modifying the in-channel number of the first convolution layer into 2. We denote the extracted optical flow features from the flow encoder as $\mathcal{F}_{fi} \in \mathbb{R}^{h_o \times w_o \times C_o}$. And for the audio encoder, we adopt the VGGish [25] model pretrained on the AudioSet [20] dataset and pool the features into a vector embedding. The obtained audio features are denoted as $\mathcal{F}_a \in \mathbb{R}^{C_a}$.

Before performing audio-visual fusion, we first employ a refinement network to fuse the extracted multi-scale image features and flow features. We first utilize the upsampling operations to ensure the refined flow features align with the visual features. Then we fuse the refined optical flow features with the visual features of scale with summation. This process can be formulated as:

$$\mathcal{F}_{refine_i} = \mathcal{F}_{\text{vi}} + \Phi_{\text{Upsample}}(\Phi_{\text{Refine}}(\mathcal{F}_{fi})), \quad (2)$$

where Φ_{Refine} is composed of multiple convolution layers.

After fusing the image and optical flow features, we obtain the aggregated visual features $\mathcal{F}_{refine_i}$; then we perform the multi-modal feature fusion and mask decoding to obtain the segmentation predictions, by inheriting the modules from TPAVI [81] or AVSegFormer [19].

4.3. Distant Frame Exploitation

Distant frames (DFs) refer to the video frames which are temporally faraway from the target labeled frame. Contrary to neighboring frames, the distant frames do not contain the motion dynamics of the target sounding objects due to the long temporal distance. However, thanks to the large visual appearance variations to the target labeled frame, these distant frames are the natural *augmentations* to the target labeled frames with shared semantics. And these frames can substantially enhance the data diversity if utilized for model training, thereby boosting the model generalization. Even though there are no groundtruth annotations for the DFs, modern pseudo-labeling techniques can be harnessed to provide self-supervision. Given the observation, we propose to exploit the unlabeled distant frames with a teacher-student network to train the model, inspired by recent works [51, 52, 56, 65, 67].

Specifically, given the training data, we first divide it into the labeled frame set and the unlabeled distant frame set. We first use the labeled frames to train the teacher for some iterations to ensure that the model has the capability to generate reliable pseudo mask labels; this step is called burn-in stage [51, 52, 56, 67]. Then we initialize the student network with the weights of the teacher; and during the training stage, the teacher and student network share the weights. The difference is that the model parameters are optimized through the student network with the supervised loss and unsupervised loss; while the teacher network with `stop_gradient` serves to provide pseudo labels for the unlabeled frames. To this end, in each iteration after the burn-in stage, the student network is optimized with labeled frames $\{(I_\ell, O_\ell, A_\ell), y_\ell\}$ and the unlabeled frames $\{(I_u, O_u, A_u)\}$. For the labeled frames, the supervised loss $\mathcal{L}_{\text{sup}}$ is computed between student prediction and groundtruth mask. For the supervised loss $\mathcal{L}_{\text{sup}}$, it can be either BCE loss [81] in TPAVI or Dice loss [57] in AVSegFormer [19], which are formulated as:

$$\mathcal{L}_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \cdot \log(p_i) + (1-y_i)\cdot \log(1-p_i)\right], \tag{3}$$

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2\sum_{i=1}^{N}(p_i \cdot y_i)}{\sum_{i=1}^{N} p_i + \sum_{i=1}^{N} y_i}, \tag{4}$$

where in Eq. (3) and Eq. (4), y_i represents the ground truth label of a pixel and p_i represents the predicted probability of a pixel belonging to the foreground class.

For the unlabeled frames $\{(I_u, O_u, A_u)\}$, the input visual signals are perturbed by two operators, i.e., weak perturbation $\mathcal{H}^w$ (*e.g.*, flip) and strong perturbation $\mathcal{H}^s$ (*e.g.*, cutmix). Afterwards, we feed the weakly-augmented view $\{\mathcal{H}^w(I_u), \mathcal{H}^w(O_u), A_u\}$ to the teacher model to predict pseudo mask label p^w; and we feed the strongly-augmented view $\{\mathcal{H}^s(I_u), \mathcal{H}^s(O_u), A_u\}$ to the student model to predict the mask p^s. The unsupervised loss $\mathcal{L}_{\text{unsup}}$ ensures that the predictions under strong perturbations align with those under weak perturbations, which can be formulated as:

$$\mathcal{L}_{\text{unsup}} = \frac{1}{B_u}\sum H(p^w, p^s), \tag{5}$$

where B_u is the batch size for unlabeled data. H serves to minimize the entropy between two probability distributions:

$$\begin{aligned} p^w &= \Phi_{\text{mask}}(H^w(I_i), H^w(O_i), A_i), \\ p^s &= \Phi_{\text{mask}}(H^s(I_i), H^s(O_i), A_i). \end{aligned} \tag{6}$$

The overall training objective $\mathcal{L}_{\text{total}}$ is a combination of supervised loss $\mathcal{L}_{sup}$ and unsupervised loss $\mathcal{L}_{\text{unsup}}$ as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{sup}} + \lambda \mathcal{L}_{\text{unsup}}, \tag{7}$$

where λ is the weight to balance the losses.

Algorithm 1 Algorithm of Our Framework

```
Input: Labeled frames $\mathcal{D}^L = \{(\mathcal{I}_\ell, \mathcal{A}_\ell), \mathcal{Y}_\ell\}$ and its unlabeled neighboring frames $\mathcal{D}_n^L = \{\mathcal{I}_n^\ell\}$, unlabeled distant frames $\mathcal{D}^D = \{\mathcal{I}_d, \mathcal{A}_d\}$ and its unlabeled neighboring frames $\mathcal{D}_n^D = \{\mathcal{I}_n^d\}$, burn-in iteration $k$, maximum iteration $N$
Output: Teacher (Student) Model Weights $\theta^i$
for $i < N$ do
    Sample labeled data from $\mathcal{D}^L$ and its NFs from $\mathcal{D}_n^L$
    Calculate the motion flow $\{O_\ell\}$
    Compute $L_{sup}$ with Eq. (3) or Eq. (4)
    if $i < k$ then
        Update $\theta^i$ with $\mathcal{L}_{sup}$
    end if
    if $i >= k$ then
        Sample unlabeled data from $\mathcal{D}^D$ and its NFs from $\mathcal{D}_n^D$
        Calculate the motion flow $\{O_d\}$
        Compute $\mathcal{L}_{unsup}$ with Eq. (5)
        Update $\theta^i$ with $\mathcal{L}_{total}$ by Eq. (7)
    end if
end for
return $\theta^i$
```

5. Experiment

5.1. Experimental setup

Datasets. We use the AVSBench [81] dataset, which was recently proposed for audio-visual segmentation task with segmentation mask annotations for sounding objects. This dataset has two subsets: the semi-supervised Single Sound Source Segmentation (S4) and the fully supervised Multiple Sound Source Segmentation (MS3). In S4 subset, there exists only one sounding object in the video. For each training video sample, only the first frame of the frame sequence is annotated while all five sampled frames need to be segmented in the validation and test sets. In MS3 subset, there might exist more than one sounding objects in the video frames. All five frames sampled from a 5s-long video are provided with mask annotations in both training and evaluation stages. In terms of the dataset size, S4 subset includes 3452/740/740 videos in training, validation and test sets separately, for a total of 10,852 annotated frames; MS3 subset includes 296/64/64 videos in training, validation and test sets, with 2,120 annotated frames. For the MS3 dataset, since all five frames are annotated in training set, we extract semantically relevant videos from the VGGSound dataset and select the middle frames as the source of distant frames for the MS3 dataset, totaling 12,990 in size.

Metrics. We adopt the mean Intersection-over-Union ($\mathcal{M}_\mathcal{J}$) and F-score ($\mathcal{M}_\mathcal{F}$) as our evaluation metrics following previous methods [19, 81].

Implementation details. Technically, our proposed framework can be combined with any mainstream methods. In our experiments, we verify our method based on TPAVI [81] and AVSegFormer [19], thus we follow their experimental settings such as backbone, learning rate and optimizing strategy. The input image and optical flow size is 224×224. For the teacher-student network, the weak augmentations include resize, crop and horizontal flip; and the strong augmentations include additional color jitter, grayscale and cutmix operations. And the loss weight λ is set to 0.5. The burn-in stage lasts for 10 epochs. We train the models for 120 epochs, with one NVIDIA A100 GPU. Batch size is 24.

5.2. Comparison with Prior Arts

Improvement of our method over baselines. To verify the effectiveness of our method, we choose two typical baseline methods: FCN-based TPAVI [81] and transformer-based AVSegFormer [19], and we apply our framework onto the baseline methods. We compare the performance between the models (**Ours**) and the baseline models in Tab. 1.

As Tab. 1 shows, our method consistently improves the performance significantly on both TPAVI [81] and AVSegFormer [19], which indicates the effectiveness and universality of our method. For the TPAVI with ResNet baseline method, our method has significant performance gains across all metrics on both subsets. On S4 subset, our model achieves 5.36 $\mathcal{M}_\mathcal{J}$ (mIoU) gains and reaches 0.887 $\mathcal{M}_\mathcal{F}$, which advances the baseline model with 0.848 $\mathcal{M}_\mathcal{F}$ by a large margin. On the more challenging MS3 subset, our model achieves higher gains of 6.20 on $\mathcal{M}_\mathcal{J}$. As for the TPAVI with PVT, although the performance of the baseline is very strong, our method can also bring performance gains. For instance, our method achieves 3.75 $\mathcal{M}_\mathcal{J}$ gains on S4 subset and 5.49 $\mathcal{M}_\mathcal{J}$ gains on MS3 subset. In terms of the AVSegFormer baseline method, even it is already a very powerful method, our method can also improve the performance upon it on both backbones. For the ResNet backbone, our method achieves 2.51 $\mathcal{M}_\mathcal{J}$ gains on S4 subset and 6.35 $\mathcal{M}_\mathcal{J}$ gains on MS3 subset. With PVT backbone, our method achieves new state-of-the-art performance: 83.15 $\mathcal{M}_\mathcal{J}$ on S4 subset and 61.95 $\mathcal{M}_\mathcal{J}$ on MS3 subset.

Method	S4		MS3	
	$\mathcal{M}_\mathcal{J}$	$\mathcal{M}_\mathcal{F}$	$\mathcal{M}_\mathcal{J}$	$\mathcal{M}_\mathcal{F}$
TPAVI$_{\text{(ResNet)}}$	72.79	.848	47.88	.578
Ours$_{\text{(ResNet)}}$	**78.15** (+5.36)	**.887**	**54.08** (+6.20)	**.616**
TPAVI$_{\text{(PVT)}}$	78.74	.879	54.00	.645
Ours$_{\text{(PVT)}}$	**82.49** (+3.75)	**.912**	**59.49** (+5.49)	**.676**
AVSegFormer$_{\text{(ResNet)}}$	76.45	.859	49.53	.628
Ours$_{\text{(ResNet)}}$	**78.96** (+2.51)	**.875**	**55.88** (+6.35)	**.645**
AVSegFormer$_{\text{(PVT)}}$	82.06	.899	58.36	.693
Ours$_{\text{(PVT)}}$	**83.15** (+1.09)	**.904**	**61.95** (+3.59)	**.709**

Table 1. Comparison of our method and baseline methods on AVSBench S4 and MS3 subsets .

Comparison with Other Arts. We also collect up-to-date AVS methods AVSC [47], CATR [44], AuTR [49], CMVAE [54], and SAMA-AVS [50]; and we compare the performance of these methods with our proposed method. The results are shown in Tab. 2. The comparison shows the strong competitiveness of our proposed framework when compared with so various latest methods. Our method based on AVSegFormer [19] is still the state-of-the-art method among all the methods. Moreover, on S4 subset, the original TPAVI method only has 72.8 $\mathcal{M}_\mathcal{J}$ with ResNet, 78.7 $\mathcal{M}_\mathcal{J}$ with PVT, which falls behind the other methods including AVSC [47], CATR [44], AuTR [49], ECMVAE [54]. However, by combining our method with the TPAVI, the model (Ours w/TPAVI) has outperformed the other methods including AVSegFormer, AVSC, CATR, AuTR and ECMVAE in almost all metrics; and it becomes the second best model except our AVSegFormer-based model "ours w/ AVSegFormer" under the same backbones. The results clearly reveal the effectiveness of our versatile proposed framework. Notably, our framework can also be applied on these methods [44, 47, 49, 50, 54] to further improve their performance.

Method	I.B.	S4		MS3	
		$\mathcal{M}_\mathcal{J}$	$\mathcal{M}_\mathcal{F}$	$\mathcal{M}_\mathcal{J}$	$\mathcal{M}_\mathcal{F}$
TPAVI [81]	ResNet	72.80	.848	47.90	.578
(ECCV'22)	PVT	78.70	.879	54.00	.645
ECMVAE [54]	ResNet	76.33	.865	48.69	.607
(ICCV'23)	PVT	81.74	.901	57.84	.708
CATR [44]	ResNet	74.80	.866	52.80	.653
(ACMMM'23)	PVT	81.40	.896	59.00	.700
AVSC [47]	ResNet	77.02	.852	49.58	.615
(ACMMM'23)	PVT	80.57	.882	58.22	.651
AuTR [49]	ResNet	75.00	.852	49.40	.612
(arXiv'23)	PVT	80.40	.891	56.20	.672
SAMA-AVS [50] (WACV'24)	ViT-H	81.53	.886	**63.14**	.691
AVSegFormer [19]	ResNet	76.45	.859	49.53	.628
(AAAI'24)	PVT	82.06	.899	58.36	.693
Ours	ResNet	78.15	.887	54.08	.616
(w/ TPAVI)	PVT	82.49	**.912**	59.49	.676
Ours	ResNet	78.96	.875	55.88	.645
(w/ AVSegFormer)	PVT	**83.15**	.904	61.95	**.709**

Table 2. Comparison with up-to-date state-of-the-arts on both subsets. Our proposed methods significantly improve the competitiveness of the baseline models. (The best performance in **bold** and the second best is underlined; "I.B." denotes image backbone.)

Method	I.B.	5%		10%	
		$\mathcal{M}_\mathcal{J}$	$\mathcal{M}_\mathcal{F}$	$\mathcal{M}_\mathcal{J}$	$\mathcal{M}_\mathcal{F}$
TPAVI	ResNet	57.34	.733	57.91	.746
	PVT	67.06	.794	71.72	.830
Ours	ResNet	61.74	.770	66.67	.820
(w/ TPAVI)	PVT	72.96	.848	76.23	.871
AVSegFormer	ResNet	56.13	.703	62.47	.754
	PVT	68.35	.797	73.92	.840
Ours	ResNet	64.96	.766	69.26	.796
(w/ AVSegFormer)	PVT	75.38	.846	77.40	.864

Table 3. Resutls of the models with different percentages of labeled training data from AVSBench S4 dataset.

Results using less labeled training data. We also investigate the performance when utilizing training data with varying proportions (5% and 10%) of labeled data on the S4 subset. As shown in Tab. 3, our approach consistently demonstrates impressive improvements across different data proportions. For instance, using only 10% of labeled data with ResNet backbone, the performance of our model increases by nearly 10 points on $\mathcal{M}_\mathcal{J}$. This indicates that our method is also highly effective when the labeled data is limited.

Method	From Scratch		P.T. on S4	
	ResNet	PVT	ResNet	PVT
TPAVI	47.90	54.00	54.30	57.30
Ours (w/ TPAVI)	**54.08**	**59.49**	**54.73**	**60.78**
AVSegFormer	49.53	58.36	55.78	61.91
Ours (w/ AVSegFormer)	**55.88**	**61.95**	**59.32**	**64.47**

Table 4. Performance with different initialization strategies under the MS3 setting on $\mathcal{M}_\mathcal{J}$.

NF	DF	$\mathcal{M}_\mathcal{J}$	$\mathcal{M}_\mathcal{F}$
✗	✗	72.80	.848
✓	✗	76.17	.869
✗	✓	77.43	.881
✓	✓	78.15	.887

(a) Results with ResNet backbone.

NF	DF	$\mathcal{M}_\mathcal{J}$	$\mathcal{M}_\mathcal{F}$
✗	✗	77.89	.880
✓	✗	81.01	.903
✗	✓	81.41	.905
✓	✓	82.49	.912

(b) Results with PVT backbone.

Table 5. Ablation study of our framework based on TPAVI baseline model on S4 subset.

Burn-in	5	10	20	30
$\mathcal{M}_\mathcal{J}$	77.72	**78.15**	77.50	77.48
$\mathcal{M}_\mathcal{F}$	0.883	**0.887**	0.881	0.882

(a) Burn-in epochs for training.

λ	0.1	0.2	0.5	1.0
$\mathcal{M}_\mathcal{J}$	77.00	77.62	**78.15**	77.19
$\mathcal{M}_\mathcal{F}$	0.878	0.883	**0.887**	0.881

(b) Unsupervised loss weight λ.

Table 6. Effects of burn-in epochs and unsupervised loss weight.

Pre-training on the Single-source subset. Following the TPAVI [81], we conduct an investigation into the impact on the MS3 when using pretrained weights of S4. As shown in Table 4, it is evident that pretraining on the S4 dataset will indeed improve the performance on the MS3 subset for all methods including ours; and our method achieves the best performance on the MS3 subset among all methods using S4-pretrained weights.

5.3. Ablation Study

In this section, we conduct ablation studies to evaluate the components in the framework with TPAVI baseline model on the S4 subset.

Effectiveness of NF and DF. To study the effects of neighboring frames (NFs) and distant frames (DFs) in our proposed framework, we conduct ablation studies, adjusting one component at a time based on the TPAVI baseline methods with two visual backbones. As illustrated in Tab. 5, both NF and DF demonstrate significant enhancements when applied independently, indicating that their respective contributions to the performance are both fairly considerable. And when combining the NF and DF, the model obtains the best performance across all metrics. The results demonstrate the effectiveness and complementarity of neighboring frames (NFs) and distant frames (DFs) in boosting the performance on the AVS task, implying the great value of the abundant unlabeled frames with proper exploitation.

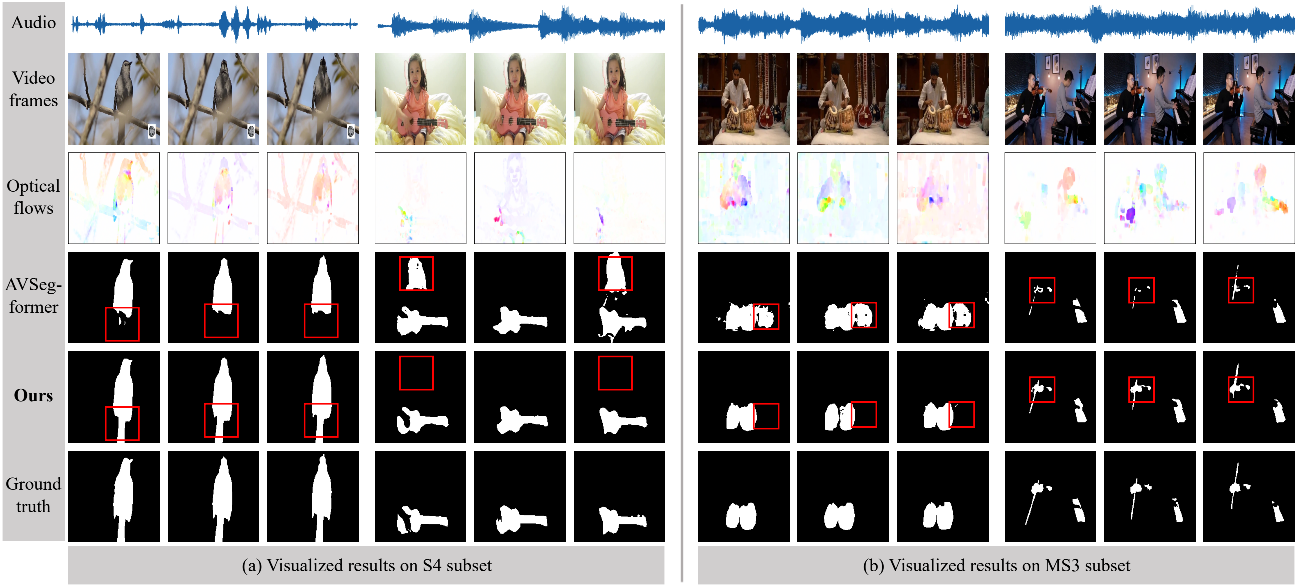

Audio
Video frames
Optical flows
AVSeg-former
Ours
Ground truth
(a) Visualized results on S4 subset
(b) Visualized results on MS3 subset

Figure 3. Qualitative comparison between our method and AVSegFormer [19] on both subsets of AVSBench. Our method shows better segmentation performance by localising the exact sounding object, attending to the fine-grained details and being closer to groundtruths.

Epochs for burn-in stage. We ablate the burn-in epochs in the teacher-student training. As shown in Tab. 6 (a), 10-epoch burn-in works best. Either using the unsupervised loss too early or too late will result in a suboptimal performance. If using the unsupervised loss too early, the pseudo labels from the teacher are not reliable thus it will cause negative effects on the final model performance. If using the unsupervised loss is too late, the model will be biased toward the labeled data without utilizing the unlabeled data.

Weight for the unsupervised loss. We ablate the weight λ for the unsupervised loss $\mathcal{L}_{\text{unsup}}$. The results in Tab. 6 (b) show a moderate value of 0.5 achieving the best performance. When λ is set as low as 0.1, the improvement is less significant than the cases where λ is 0.2 or 0.5. However, if λ is too high such as 1.0, the model performance degrades. This is due to the model overemphasizing the unlabeled data when using large unsupervised loss weight.

5.4. Qualitative Examples

In Fig. 3, we qualitatively show some segmentation results of our method and the baseline AVSegFormer [19]. The results clearly demonstrate the advantages of our method by producing the segmentations for the sounding objects which are closer to groundtruths. As shown in the first video of Fig. 3 (a), with the assistance of flow, our method can segment the tail of the bird which is hard to find with only visual RGB frames. In the second video of Fig. 3 (a), the AVSegFormer baseline falsely segments both the *Ukulele* and the salient yet silent *person*. On the contrary, our method accurately localizes and segments only *Ukulele* according to the audio cues, without being distracted by the silent *person*. In the multi-sound scenario in Fig. 3 (b), our method also shows improved performance. In the first video, the AVSegFormer [19] baseline mistakenly segments another silent instrument as marked with red boxes; while our method leverages both the sound and the hand motion of the player to produce the segmentations for the instrument being played. In the second video, the segmentations produced by our method for both instruments have fine-grained details and are closer to groundtruth annotations.

6. Conclusion

In this paper, we have pointed out the major limitation of previous AVS methods: the *underutilization* of the abundant unlabeled frames. To mitigate this, we analyzed that the unlabeled frames can be divided into two categories: *neighboring frame* (NF) and *distant frame* (DF), according to the temporal characteristics. And we proposed a unified *unlabeled frame exploitation* (UFE) framework to harness the two kinds of unlabeled frames based on their unique traits. For NFs, we extracted the motion cues as dynamic guidance to assist in the precise localization of sounding objects; while for DFs, since they are natural semantic augmentations to the labeled frames, we utilized them to enrich the data diversity with the teacher-student training. Extensive experiments have demonstrated the significant improvement brought by the exploitation of unlabeled frames. We believe that our proposed framework serves as a strong baseline and hopefully inspires more research to value both labeled and unlabeled data to successfully tackle AVS task.

References

[1] Madhav Agarwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. Audio-visual face reenactment. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5178–5187, 2023. 1

[2] Zohar Barzelay and Yoav Y. Schechner. Harmony in motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007. 2, 3

[3] 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*, pages 16867–16876, 2021. 2

[4] Zida Cheng, Shuai Xiao, Zhonghua Zhai, Xiaoyi Zeng, and Weilin Huang. Mixer: Image to multi-modal retrieval learning for industrial application. *arXiv preprint arXiv:2305.03972*, 2023. 2

[5] Joon Son Chung, Andrew W. Senior, Oriol Vinyals, and Andrew Zisserman. Lip reading sentences in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3444–3453. IEEE Computer Society, 2017. 3

[6] Marco Cristani, Manuele Bicego, and Vittorio Murino. Audio-visual event recognition in surveillance video sequences. *IEEE Transactions on Multimedia*, 9(2):257–267, 2007. 1

[7] Marco Crocco, Marco Cristani, Andrea Trucco, and Vittorio Murino. Audio surveillance: A systematic review. *ACM Computing Surveys (CSUR)*, 48(4):1–46, 2016. 1

[8] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T. Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation. *ACM Trans. Graph.*, 37(4):112, 2018. 2, 3

[9] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T. Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation. *ACM Trans. Graph.*, 37(4):112, 2018. 2

[10] Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Image Analysis: 13th Scandinavian Conference, SCIA 2003 Halmstad, Sweden, June 29–July 2, 2003 Proceedings 13*, pages 363–370. Springer, 2003. 4

[11] Dennis Fedorishin, Deen Dayal Mohan, Bhavin Jawade, Srirangaraj Setlur, and Venu Govindaraju. Hear the flow: Optical flow-based self-supervised visual sound source localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2278–2287, 2023. 3

[12] John W Fisher III, Trevor Darrell, William Freeman, and Paul Viola. Learning joint statistical models for audio-visual fusion and segregation. *Advances in Neural Information Processing Systems*, 13, 2000. 2, 3

[13] Denis Fortun, Patrick Bouthemy, and Charles Kervrann. Optical flow modeling and computation: A survey. *Computer Vision and Image Understanding*, 134:1–21, 2015. 4

[14] Jules Françoise, Norbert Schnell, Riccardo Borghesi, and Frédéric Bevilacqua. Probabilistic models for designing motion and sound relationships. In *Proceedings of the 2014 international conference on new interfaces for musical expression*, pages 287–292, 2014. 3

[15] Geoffrey French, Samuli Laine, Timo Aila, Michal Mackiewicz, and Graham D. Finlayson. Semi-supervised semantic segmentation needs strong, varied perturbations. In *Proceedings of the British Machine Vision Conference*. BMVA Press, 2020. 3

[16] Aviv Gabbay, Ariel Ephrat, Tavi Halperin, and Shmuel Peleg. Seeing through noise: Speaker separation and enhancement using visually-derived speech. *CoRR*, abs/1708.06767, 2017. 3

[17] Chuang Gan, Yiwei Zhang, Jiajun Wu, Boqing Gong, and Joshua B Tenenbaum. Look, listen, and act: Towards audio-visual embodied navigation. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9701–9707. IEEE, 2020. 1

[18] 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*, pages 15490–15500. IEEE, 2021. 2

[19] Shengyi Gao, Zhe Chen, Guo Chen, Wenhai Wang, and Tong Lu. Avsegformer: Audio-visual segmentation with transformer. *CoRR*, abs/2307.01146, 2023. 1, 2, 3, 4, 5, 6, 7, 8

[20] 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 *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017. 4

[21] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129:1789–1819, 2021. 3

[22] Longyin Guo, Qijun Zhao, and Hongmei Gao. Look longer to see better: Audio-visual event localization by exploiting long-term correlation. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 01–07. IEEE, 2022. 2

[23] Sharon E Guttman, Lee A Gilroy, and Randolph Blake. Hearing what the eyes see: Auditory encoding of visual temporal sequences. *Psychological science*, 16(3):228–235, 2005. 1, 2

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 4

[25] 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 *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 131–135. IEEE, 2017. 4

[26] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean.

Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015. 3

[27] Di Hu, Feiping Nie, and Xuelong Li. Deep multimodal clustering for unsupervised audiovisual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9248–9257, 2019. 2

[28] Di Hu, Rui Qian, Minyue Jiang, Xiao Tan, Shilei Wen, Errui Ding, Weiyao Lin, and Dejing Dou. Discriminative sounding objects localization via self-supervised audiovisual matching. *Advances in Neural Information Processing Systems*, 33:10077–10087, 2020.

[29] Di Hu, Yake Wei, Rui Qian, Weiyao Lin, Ruihua Song, and Ji-Rong Wen. Class-aware sounding objects localization via audiovisual correspondence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2

[30] Hanzhe Hu, Fangyun Wei, Han Hu, Qiwei Ye, Jinshi Cui, and Liwei Wang. Semi-supervised semantic segmentation via adaptive equalization learning. *Advances in Neural Information Processing Systems*, 34:22106–22118, 2021. 3

[31] Chen Ju, Peisen Zhao, Ya Zhang, Yanfeng Wang, and Qi Tian. Point-level temporal action localization: Bridging fully-supervised proposals to weakly-supervised losses. *arXiv preprint arXiv:2012.08236*, 2020. 3

[32] Chen Ju, Peisen Zhao, Siheng Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. Divide and conquer for single-frame temporal action localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13455–13464, 2021. 3

[33] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. In *European Conference on Computer Vision*, pages 105–124. Springer, 2022. 2

[34] Chen Ju, Peisen Zhao, Siheng Chen, Ya Zhang, Xiaoyun Zhang, Yanfeng Wang, and Qi Tian. Adaptive mutual supervision for weakly-supervised temporal action localization. *IEEE Transactions on Multimedia*, 2022. 3

[35] Chen Ju, Zeqian Li, Peisen Zhao, Ya Zhang, Xiaopeng Zhang, Qi Tian, Yanfeng Wang, and Weidi Xie. Multi-modal prompting for low-shot temporal action localization. *arXiv preprint arXiv:2303.11732*, 2023. 2

[36] Chen Ju, Haicheng Wang, Zeqian Li, Xu Chen, Zhonghua Zhai, Weilin Huang, and Shuai Xiao. Turbo: Informativity-driven acceleration plug-in for vision-language models. *arXiv preprint arXiv:2312.07408*, 2023.

[37] Chen Ju, Haicheng Wang, Jinxiang Liu, Chaofan Ma, Ya Zhang, Peisen Zhao, Jianlong Chang, and Qi Tian. Constraint and union for partially-supervised temporal sentence grounding. *arXiv preprint arXiv:2302.09850*, 2023. 2

[38] Chen Ju, Kunhao Zheng, Jinxiang Liu, Peisen Zhao, Ya Zhang, Jianlong Chang, Qi Tian, and Yanfeng Wang. Distilling vision-language pre-training to collaborate with weakly-supervised temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14751–14762, 2023. 3

[39] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5492–5501, 2019. 2

[40] E. Kidron, Y.Y. Schechner, and M. Elad. Pixels that sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 88–95 vol. 1, 2005. 3

[41] Jun-Tae Lee, Mihir Jain, Hyoungwoo Park, and Sungrack Yun. Cross-attentional audio-visual fusion for weakly-supervised action localization. In *International Conference on Learning Representations*, 2020. 2

[42] Seung Hyun Lee, Gyeongrok Oh, Wonmin Byeon, Chanyoung Kim, Won Jeong Ryoo, Sang Ho Yoon, Hyunjun Cho, Jihyun Bae, Jinkyu Kim, and Sangpil Kim. Sound-guided semantic video generation. In *European Conference on Computer Vision*, pages 34–50. Springer, 2022. 1

[43] Seung Hyun Lee, Wonseok Roh, Wonmin Byeon, Sang Ho Yoon, Chanyoung Kim, Jinkyu Kim, and Sangpil Kim. Sound-guided semantic image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3377–3386, 2022. 1

[44] Kexin Li, Zongxin Yang, Lei Chen, Yi Yang, and Jun Xiao. Catr: Combinatorial-dependence audio-queried transformer for audio-visual video segmentation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1485–1494, 2023. 1, 2, 3, 6, 7

[45] Yan-Bo Lin, Hung-Yu Tseng, Hsin-Ying Lee, Yen-Yu Lin, and Ming-Hsuan Yang. Unsupervised sound localization via iterative contrastive learning. *CoRR*, abs/2104.00315, 2021. 2

[46] Yan-Bo Lin, Yu-Jhe Li, and Yu-Chiang Frank Wang. Dual-modality seq2seq network for audio-visual event localization. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2002–2006. IEEE, 2019. 2

[47] Chen Liu, Peike Patrick Li, Xingqun Qi, Hu Zhang, Lincheng Li, Dadong Wang, and Xin Yu. Audio-visual segmentation by exploring cross-modal mutual semantics. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7590–7598, 2023. 1, 2, 3, 6, 7

[48] Jinxiang Liu, Chen Ju, Weidi Xie, and Ya Zhang. Exploiting transformation invariance and equivariance for self-supervised sound localisation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3742–3753, 2022. 2

[49] Jinxiang Liu, Chen Ju, Chaofan Ma, Yanfeng Wang, Yu Wang, and Ya Zhang. Audio-aware query-enhanced transformer for audio-visual segmentation. *CoRR*, abs/2307.13236, 2023. 1, 2, 3, 6, 7

[50] Jinxiang Liu, Yu Wang, Chen Ju, Chaofan Ma, Ya Zhang, and Weidi Xie. Annotation-free audio-visual segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5604–5614, 2024. 1, 2, 3, 6, 7

[51] Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Peter Vajda. Unbiased teacher for semi-supervised object detection. In *Proceedings of the International Conference on Learning Representations*. OpenReview.net, 2021. 3, 5

[52] Yen-Cheng Liu, Chih-Yao Ma, and Zsolt Kira. Unbiased teacher v2: Semi-supervised object detection for anchor-free and anchor-based detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9819–9828, 2022. 5

[53] Tanvir Mahmud and Diana Marculescu. Ave-clip: Audioclip-based multi-window temporal transformer for audio visual event localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5158–5167, 2023. 2

[54] Yuxin Mao, Jing Zhang, Mochu Xiang, Yiran Zhong, and Yuchao Dai. Multimodal variational auto-encoder based audio-visual segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 954–965, 2023. 1, 2, 3, 6, 7

[55] Luke Melas-Kyriazi and Arjun K Manrai. Pixmatch: Unsupervised domain adaptation via pixelwise consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12435–12445, 2021. 3

[56] Peng Mi, Jianghang Lin, Yiyi Zhou, Yunhang Shen, Gen Luo, Xiaoshuai Sun, Liujuan Cao, Rongrong Fu, Qiang Xu, and Rongrong Ji. Active teacher for semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14482–14491, 2022. 5

[57] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. IEEE, 2016. 5

[58] Andrew Owens and Alexei A. Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VI*, pages 639–658. Springer, 2018. 3

[59] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision*, pages 631–648, 2018. 2

[60] Rui Qian, Di Hu, Heinrich Dinkel, Mengyue Wu, Ning Xu, and Weiyao Lin. Multiple sound sources localization from coarse to fine. In *Proceedings of the European Conference on Computer Vision*, pages 292–308. Springer, 2020. 2

[61] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 4358–4366, 2018. 2

[62] Mykhailo Shvets, Wei Liu, and Alexander C Berg. Leveraging long-range temporal relationships between proposals for video object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9756–9764, 2019. 2

[63] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in Neural Information Processing Systems*, 27, 2014. 4

[64] Rajsuryan Singh, Pablo Zinemanas, Xavier Serra, Juan Pablo Bello, and Magdalena Fuentes. Flowgrad: Using motion for visual sound source localization. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 2, 3

[65] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in Neural Information Processing Systems*, 33:596–608, 2020. 3, 5

[66] Zengjie Song, Yuxi Wang, Junsong Fan, Tieniu Tan, and Zhaoxiang Zhang. Self-supervised predictive learning: A negative-free method for sound source localization in visual scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3222–3231, 2022. 2

[67] Jiamu Sun, Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Guannan Jiang, Zhiyu Wang, and Rongrong Ji. Refteacher: A strong baseline for semi-supervised referring expression comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19144–19154, 2023. 5

[68] Yihe Tang, Weifeng Chen, Yijun Luo, and Yuting Zhang. Humble teachers teach better students for semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3132–3141, 2021. 3

[69] Yapeng Tian, Dingzeyu Li, and Chenliang Xu. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In *Proceedings of the European Conference on Computer Vision*, pages 436–454. Springer, 2020. 2

[70] Efthymios Tzinis, Scott Wisdom, Tal Remez, and John R. Hershey. Audioscopev2: Audio-visual attention architectures for calibrated open-domain on-screen sound separation. In *Proceedings of the European Conference on Computer Vision*, pages 368–385. Springer, 2022. 2

[71] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2, 3

[72] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424, 2022. 4

[73] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018. 2

[74] Xinyu Wu, Haitao Gong, Pei Chen, Zhi Zhong, and Yangsheng Xu. Surveillance robot utilizing video and audio information. *Journal of Intelligent and Robotic Systems*, 55: 403–421, 2009. 1

[75] Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3060–3069, 2021. 3

[76] Lihe Yang, Lei Qi, Litong Feng, Wayne Zhang, and Yinghuan Shi. Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7236–7246, 2023. 3

[77] 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*, pages 570–586, 2018. 2

[78] 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*, pages 1735–1744. IEEE, 2019. 2, 3

[79] 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*, pages 1735–1744, 2019. 2

[80] Peisen Zhao, Lingxi Xie, Chen Ju, Ya Zhang, Yanfeng Wang, and Qi Tian. Bottom-up temporal action localization with mutual regularization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, pages 539–555. Springer, 2020. 3

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

[82] Yuliang Zou, Zizhao Zhang, Han Zhang, Chun-Liang Li, Xiao Bian, Jia-Bin Huang, and Tomas Pfister. Pseudoseg: Designing pseudo labels for semantic segmentation. In *Proceedings of the International Conference on Learning Representations*. OpenReview.net, 2021. 3