Egocentric Audio-Visual Object Localization

Chao Huang[1], Yapeng Tian[1], Anurag Kumar[2], Chenliang Xu[1]
[1]University of Rochester, [2]Meta Reality Labs Research
{chaohuang, yapengtian, chenliang.xu}@rochester.edu, anuragkr90@fb.com

Abstract

Humans naturally perceive surrounding scenes by unifying sound and sight from a first-person view. Likewise, machines are advanced to approach human intelligence by learning with multisensory inputs from an egocentric perspective. In this paper, we explore the challenging egocentric audio-visual object localization task and observe that 1) egomotion commonly exists in first-person recordings, even within a short duration; 2) The out-of-view sound components can be created when wearers shift their attention. To address the first problem, we propose a geometry-aware temporal aggregation module that handles the egomotion explicitly. The effect of egomotion is mitigated by estimating the temporal geometry transformation and exploiting it to update visual representations. Moreover, we propose a cascaded feature enhancement module to overcome the second issue. It improves cross-modal localization robustness by disentangling visually-indicated audio representation. During training, we take advantage of the naturally occurring audio-visual temporal synchronization as the "free" self-supervision to avoid costly labeling. We also annotate and create the Epic Sounding Object dataset for evaluation purposes. Extensive experiments show that our method achieves state-of-the-art localization performance in egocentric videos and can be generalized to diverse audio-visual scenes. Code is available at https://github.com/WikiChao/Ego-AV-Loc.

1. Introduction

The emergence of wearable devices has drawn the attention of the research community to egocentric videos, the significance of which can be seen from egocentric research in a variety of applications such as robotics [32,34,48], augmented/virtual reality [31,61,75], and healthcare [53,66]. In recent years, the computer vision community has made substantial efforts to build benchmarks [12, 13, 15, 40, 57, 69], establish new tasks [17, 36, 37, 39, 60], and develop frameworks [33, 41, 54, 82] for egocentric video understanding.

While existing works achieve promising results in the egocentric domain, it still remains an interesting but challenging topic to perform fine-grained egocentric video understanding. For instance, understanding which object is emitting sound in a first-person recording is difficult for machines. As shown in Fig. 1, the wearer moves his/her head to put down the bottle. The frying pot which emits sound subsequently suffers deformation and occlusion due to the wearer's egomotion. Human speech outside the wearer's view also affects the machine's understanding of the current scene. This example reveals two significant challenges for designing powerful and robust egocentric video understanding systems: First, people with wearable devices usually record videos in naturalistic surroundings, where a variety of illumination conditions, object appearance, and motion patterns are shown. The dynamic visual variations introduce difficulties in accurate visual perception. Second, egocentric scenes are often perceived within a limited field of view (FoV). The common body and head movements cause frequent view changes (see Fig. 1), which brings object deformation and creates dynamic out-of-view content.

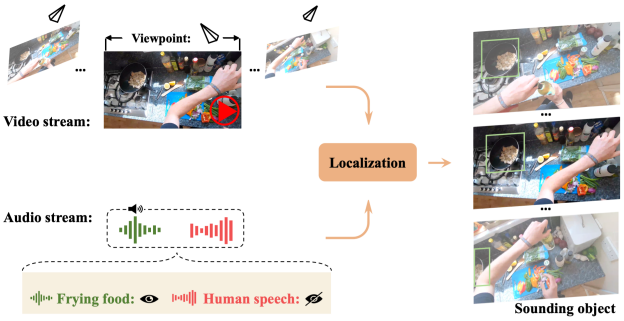


Figure 1. **Sounding object localization in egocentric videos.** Due to the wearer's egomotion, the viewpoint changes continuously across time. Consequently, audio-visual relations are dynamically changing in egocentric videos. Our approach tackles challenges in the egocentric audio-visual sounding object task and learns audio-visual associations from first-person videos.

Although a visual-only system may struggle to fully decode the surrounding information and perceive scenes in egocentric videos, audio provides stable and persistent signals associated with the depicted events. Instead of purely visual perception, numerous psychological and cognitive

studies [6, 29, 67, 73] show that integration of auditory and visual signals is significant in human perception. Audio, as an essential but less focused modality, often provides synchronized and complementary information with the video stream. In contrast to the variability of first-person visual footage, sound describes the underlying scenes consistently. These natural characteristics make audio another indispensable ingredient for egocentric video understanding.

To effectively leverage audio and visual information in egocentric videos, a pivotal problem is to analyze the fine-grained audio-visual association, specifically identifying which objects are emitting sounds in the scene. In this paper, we explore a novel egocentric audio-visual object localization task, which aims to associate audio with dynamic visual scenes and localize sounding objects in egocentric videos. Given the dynamic nature of egocentric videos, it is exceedingly challenging to link visual content from various viewpoints with audio captured from the entire space. Hence, we develop a new framework to model the distinct characteristics of egocentric videos by integrating audio. In the framework, we propose a geometry-aware temporal module to handle egomotion explicitly. Our approach mitigates the impact of egomotion by performing geometric transformations in the embedding space and aligning visual features from different frames. We further use the aligned features to leverage temporal contexts across frames to learn discriminative cues for localization. Additionally, we introduce a cascaded feature enhancement module to handle out-of-view sounds. The module helps mitigate audio noises and improves cross-modal localization robustness.

Due to the dynamic nature of egocentric videos, it is hard and costly to label sounding objects for supervised training. To avoid tedious labeling, we formulate this task in a self-supervised manner, and our framework is trained with audio-visual temporal synchronization. Since there are no publicly available egocentric sounding object localization datasets, we annotate an *Epic Sounding* dataset to facilitate research in this field. Experimental results demonstrate that modeling egomotion and mitigating out-of-view sound can improve egocentric audio-visual localization performance.

In summary, our contributions are: (1) the first systematical study on egocentric audio-visual sounding object localization; (2) an effective geometry-aware temporal aggregation approach to deal with unique egomotion; (3) a novel cascaded feature enhancement module to progressively inject localization cues; and (4) an *Epic Sounding Object* dataset with sounding object annotations to benchmark the localization performance in egocentric videos.

2. Related Work

Audio-visual learning in third-person view videos. Taking the natural audio-visual synchronization in videos, a large number of studies in the past few years have proposed to jointly learn from both auditory and visual modalities. We have seen a spectrum of new audio-visual problems and applications, including visually guided sound source separation [16, 20–23, 63, 76, 87, 88], audio-visual representation learning [2, 3, 5, 26, 35, 58, 59], audio-visual event localization [44, 78, 79, 84], audio-visual video parsing [77, 83], and sounding object visual localization [4, 11, 27, 28, 38, 51, 52, 62, 65, 71]. Most previous approaches learn audio-visual correlations from third-person videos, while the distinct challenges of audio-visual learning in egocentric videos are underexplored. Different from existing works, we propose an audio-visual learning framework to explicitly solve egomotion and out-of-view audio issues in egocentric videos.

Egocentric video understanding. In the last decade, video scene understanding techniques thrived because of the well-defined third-person video datasets [7, 9, 49, 72]. Nevertheless, most of the algorithms are developed to tackle videos curated by human photographers. The natural characteristics of egocentric video data, *e.g.*, view changes, large motions, and visual deformation, are not well-explored. To bridge this gap, multiple egocentric datasets [12, 13, 24, 36, 69, 74] have been collected. These datasets have significantly advanced investigations on egocentric video understanding problems, including activity recognition [33, 41, 89], human(hand)-object interaction [8, 14, 55, 68], anticipation [1, 19, 45, 70], and human body pose inferring [30, 56]. However, only a handful of audio-visual works [10, 33, 50, 85] is presented for egocentric video understanding. Among those, Kazakos *et al*. [33] proposed an audio-visual fusion network for action recognition, while Mittal *et al*. [50] used audible interactions as cues to learn state-aware visual representations in egocentric videos. There are limited studies in explicit egomotion mitigation and fine-grained audio-visual association learning in egocentric videos. Unlike past works, we tackle challenges in egocentric audio-visual data and propose a robust sounding object localization framework. To enable the research, we propose *Epic Sounding Object* dataset based on Epic-Kitchens [12, 13].

3. Method

Our goal is to localize sounding objects in egocentric videos visually. We start by formulating our egocentric audio-visual object localization task in Sec. 3.1. Our proposed method includes a feature extraction process (described in Sec. 3.2), a two-stage cascaded feature enhancement pipeline (in Sec. 3.3), and a geometry-aware temporal aggregation module (explained in Sec. 3.4). Finally, we summarize the overall training objective in Sec. 3.5.

3.1. Problem Formulation and Method Overview

Given an egocentric video clip $V = \{I_i\}_{i=1}^{T}$ in T frames and its corresponding sound stream s, sounding object vi-

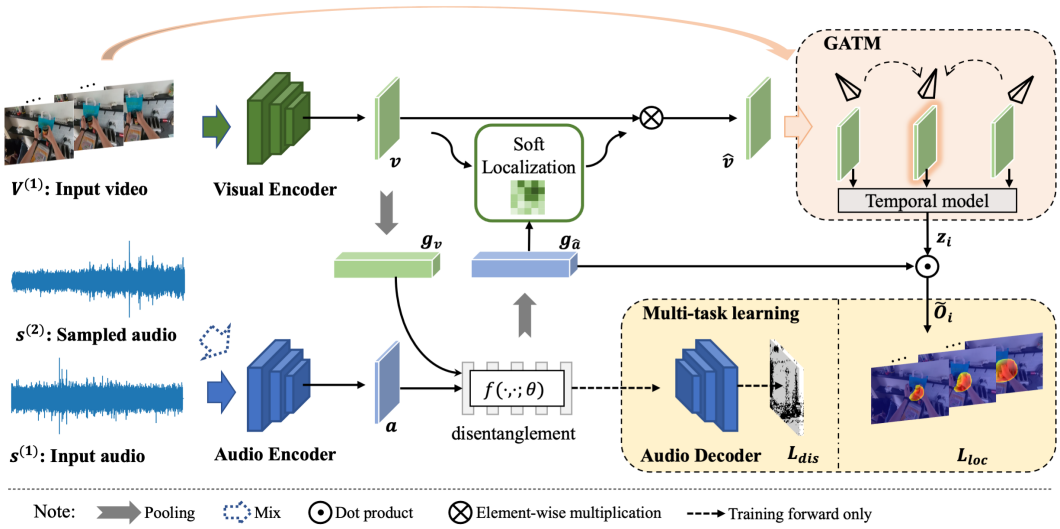


Figure 2. An overview of our egocentric audio-visual object localization framework. In the beginning, our model extracts deep features from the video and audio streams. Then, the audio and visual features are fed into the cascaded feature enhancement module to inject localization cues for both branches. Such a module is additionally trained with "mix-and-separation" strategy. Next, our geometric-aware temporal modeling block leverages the relative geometric information between visual frames and performs temporal context aggregation to get the final visual features for localization.

sual localization aims at predicting location maps $\mathcal{O} = \{O_i\}_{i=1}^{T}$ that represent sounding objects in the egocentric video. Specifically, $O_i(x, y) \in \{0, 1\}$ and positive visual regions indicate locations of sounding objects. In real-world scenarios, the captured sound can be a mixture of multiple sound sources $s = \sum_{n=1}^{N} s_n$, where s_n is the n-th sound source and it could be out of view. For the visual input, the video frames may be captured from different viewpoints. To design a robust and effective egocentric audio-visual sounding object localization system, we should consider the above issues in egocentric audio and visual data and answer two key questions: (Q1) how to associate visual content with audio representations while out-of-view sounds may exist; (Q2) how to persistently associate audio features with visual content that are captured under different viewpoints.

Due to the dynamic nature of egocentric videos, it is difficult and costly to annotate sounding objects for supervised training. To bypass the tedious labeling, we solve the egocentric audio-visual object localization task in a self-supervised manner. The proposed framework is shown in Fig. 2. Our model first extracts representations from the audio s and video clip V. In order to handle Q1, we develop a cascaded feature enhancement module to disentangle visually indicated sound sources and attend to visual regions that correspond to the visible sound sources. To enable the disentanglement, we use on-screen sound separation task as the proxy and adopt a multi-task learning objective to train our model where the localization task is solved along with a sound-separation task. To deal with the egomotion in egocentric videos (Q2), we design a geometry-aware temporal modeling approach to mitigate the feature distortion brought by viewpoint changes and aggregate the visual features temporally. We take the audio-visual temporal synchronization as the supervision signal and estimate the localization map $\tilde{O}_i$.

3.2. Feature Extraction

Visual representation. We use a visual encoder network E_v to extract visual feature maps from each input frame I_i. In our implementation, a pre-trained Dilated ResNet [86] model is adopted by removing the final fully-connected layer. We can subsequently obtain a group of feature maps $v_i = E_v(I_i)$, where $v_i \in \mathbb{R}^{c \times h_v \times w_v}$. Here c is the number of channels, and $h_v \times w_v$ denotes the spatial size.

Audio representation. To extract audio representations from the input raw waveform, we first transform audio stream s into a magnitude spectrogram X with the short-time Fourier transform (STFT). Then, we extract audio features $a = E_a(X), a \in \mathbb{R}^{c \times h_a \times w_a}$ by means of a CNN encoder E_a in the Time-Frequency (T-F) space.

3.3. Cascaded Feature Enhancement

As discussed in Sec. 3.1, a sound source s_n in the mixture s could be out of view due to constant view changes in egocentric videos and the limited FoV. This poses challenges in visually localizing sound sources and performance can degrade when the audio-visual associations are not precise. To address this, we update the features in a cascaded fashion. We first force the network to learn disentangled audio representations from the mixture using visual guidance. Then we utilize the disentangled audio representations to inject the visual features with more localization cues.

Disentanglement through sound source separation. Sound source localization objective can implicitly guide the system to learn disentangled audio features as the network will try to precisely localize the sound, and in turn, the on-screen sound will get disentangled from the rest. However, we formulate our problem in an unsupervised setting where labels for such localization objective are not available.

Audio-visual sound separation task [23, 87] uses visual information as guidance to learn to separate individual sounds from a mixture. Given the visual guidance, it is expected that the learned representations primarily encode information from visually indicated sound sources. Hence we argue for a multi-task learning approach to solve our primary task. Along with the audio-visual sounding object localization task, the network also learns to disentangle visible audio representations from the mixture through a source separation task.

- **Training.** We adopt the commonly used "mix-and-separate" strategy [23,87] for audio-visual sound separation. Given the current audio $s^{(1)}$, we randomly sample another audio stream $s^{(2)}$ from a different video and mix them together to generate input audio mixture $\tilde{s} = s^{(1)} + s^{(2)}$. We then obtain magnitude spectrograms $\tilde{X}$, $X^{(1)}$, $X^{(2)}$ for $\tilde{s}$, $s^{(1)}$ and $s^{(2)}$ respectively. The audio features is then modified as $a = E_a(\tilde{X})$.
- **Inference.** During inference, we take the original audio stream as input: $s = s^{(1)}$ and $X = X^{(1)}$ to extract visually correlated audio representations. Note that the audio features is $a = E_a(X)$.

We define the audio disentanglement network as a network $f(\cdot)$, which produces the disentangled audio features $\hat{a} \in \mathbb{R}^{c \times h_a \times w_a}$. In this network, we want to associate the visual content with the audio representations to perform disentanglement in the embedding space. Concretely, we first apply spatial average pooling on each v_i and temporal max pooling along the time axis to obtain a visual feature vector $g_v \in \mathbb{R}^c$. Then we replicate the visual feature vector $h_a \times w_a$ times and tile them to match the size of a. We concatenate the visual and audio feature maps along the channel dimension and feed them into the network. Therefore, the audio feature disentanglement can be formulated as:

$$\hat{a} = f(\textsc{Concat}[a, \textsc{Tile}(g_v)]). \tag{1}$$

In practice, we implement the disentanglement network f using two 1x1 convolution layers. The audio feature $\hat{a}$ will be used for both separation mask and sounding object localization map generation.

To separate visible sounds, we add an audio decoder D_a following the disentanglement network to output a binary mask $M_{pred} = D_a(\hat{a})$ (at the bottom of Fig. 2). U-Net architectures [23] are used in the audio encoder E_a and decoder D_a. We implement the E_a and D_a in five convolution and up-convolution layers, respectively. Details of the network architectures are provided in the Appendix. The ground truth separation mask M_{gt} can be calculated by determining whether the original input sound is dominant at locations (u, v) in the T-F space:

$$M_{gt}(u, v) = [X^{(1)}(u, v) \geq \tilde{X}(u, v)]. \tag{2}$$

To train the sound separator, we minimize the ℓ_2 distance between the predicted and ground-truth masks as the disentanglement learning objective:

$$\mathcal{L}_{dis} = ||M_{pred} - M_{gt}||_2^2. \tag{3}$$

Soft localization. Similar to the out-of-view sounds, the visual frames may contain sound-irrelevant regions. In order to learn more precise audio-visual associations, we propose to highlight the spatial regions that are more likely to be correlated with the on-screen sounds by computing audio-visual attention. The attention map will indicate the correlation between audio and visual representations at different spatial locations. Given the output $\hat{a}$ from disentanglement network $f(\cdot)$, we apply max pooling on its time and frequency dimensions, obtaining an audio feature vector $g_{\hat{a}}$. Then at each spatial position (x,y) of visual feature v_i, we compute the cosine similarity between audio and visual feature vectors:

$$S_i : S_i(x, y) = \textsc{CosineSim}(v_i(x, y), g_{\hat{a}}). \tag{4}$$

$\textsc{Softmax}$ is then used on S_i to generate a soft mask that represents the audio-visual correspondence. Hence, each v_i can be attended with the calculated weights:

$$\hat{v}_i = \textsc{Softmax}(S_i) \cdot v_i. \tag{5}$$

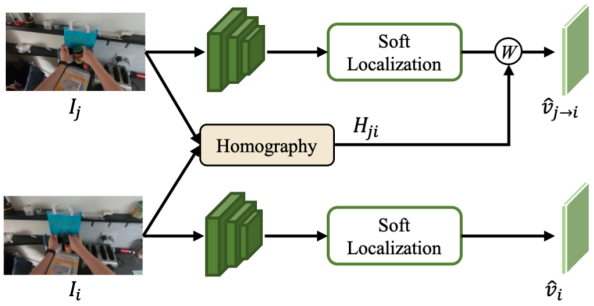


Figure 3. Overview of our proposed geometry-aware modeling approach. The visual features $\hat{v}_j$ are warped to viewpoint i by homography transformation.

3.4. Geometry-Aware Temporal Modeling

Given the temporal nature of sounds and the persistence of audio-visual associations, we incorporate temporal information from neighboring frames to learn sounding object features. However, temporal modeling is a challenging

problem for egocentric videos due to widespread egomotion and object-appearance deformations.

Although visual objects are dynamically changing, the surrounding physical environment is persistent. Hence, temporal variations in egocentric videos reveal rich 3D geometric cues that can recover the surrounding scene from changing viewpoints. Prior works have shown that given a sequence of frames, one can reconstruct the underlying 3D scene from the 2D observations [64,81]. In our work, rather than reconstructing the 3D structures, we estimate the relative geometric transformation between frames to alleviate egomotion. Specifically, we apply the transformation at the feature level to perform geometry-aware temporal aggregation. Given $\{I_i\}_{i=1}^T$ and their features $\{\hat{v}_i\}_{i=1}^T$, we take $\hat{v}_i$ as a query at a time and use the other features from neighboring frames as support features to aggregate temporal contexts. For clarity, we decompose the geometry-aware temporal aggregation into two parts: geometry modeling and temporal aggregation.

Geometry modeling. This step aims to compute the geometric transformation that represents the egomotion between frames (see Fig. 3). We found that homography estimation, which can align images taken from different perspectives, can serve as a way to measure geometric transformation. We adopt SIFT [46] + RANSAC [18] to solve homography. To be specific, a homography is a 3×3 matrix that consists of 8 degree of freedom (DOF) for scale, translation, rotation, and perspective respectively. Given the query frame I_i and a supporting frame I_j, we use $h(\cdot)$ to denote the computation process:

$$\mathcal{H}_{ji} = h(I_j, I_i)_{j \to i}, \tag{6}$$

where $\mathcal{H}_{ji}$ represents the homography transformation from frame I_j to I_i. With the computed homography transformation, we can then apply it at the feature level to transform visual features $\hat{v}_j$ to $\hat{v}_{ji}$. The $\hat{v}_{ji}$ is egomotion-free under the viewpoint of I_i. Since the resolution of feature maps is scaled down compared to the raw frame size, the homography matrix $\mathcal{H}$ should also be downsampled using the same scaling factor. The feature transformation can be written as:

$$\hat{v}_{ji} = \mathcal{H}_{ji} \otimes \hat{v}_j, \tag{7}$$

where $\otimes$ represents the warping operation.

Temporal aggregation. For the query feature $\hat{v}_i$, we end up with set of aligned features $\{\hat{v}_{ji}\}_{j=1}^T$ corresponding to each frame viewpoint. To aggregate the temporal contexts, we propose to compute the correlation between features from different frames at the same locations (see Fig. 4). The aggregation process can be formulated as:

$$z_i(x,y) = \hat{v}_i(x,y) + \textsc{Softmax}\left(\frac{\hat{v}_i(x,y)\hat{\boldsymbol{v}}(x,y)^T}{\sqrt{d}}\right)\hat{\boldsymbol{v}}(x,y), \tag{8}$$

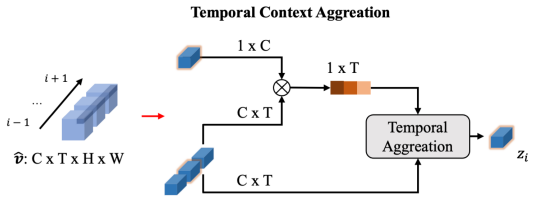


Figure 4. Illustration on the temporal context aggregation process for query feature $\hat{v}_i$ and neighboring frame features $\hat{v}_{i-1}$ and $\hat{v}_{i+1}$, which is performed independently at each spatial location.

where $\hat{\boldsymbol{v}} = [\hat{v}_{1i}; ...; \hat{v}_{Ti}]$ is the concatenation of frame features; the scaling factor d is equal to the feature dimension; and $(\cdot)^T$ represents the transpose operation [80]. The aggregation operation is applied at all spatial locations (x, y) to generate the updated visual features z_i.

3.5. Training Objective

We take audio-visual synchronization as the "free" supervision and solve the task in a self-supervised manner using contrastive learning [2, 4, 11, 65].

With the audio feature vector $g_{\hat{a}}$ and the visual features $\{z_i\}_{i=1}^T$, we can compute an audio-visual attention map S_i in Eq. 4 for each frame I_i. The training objective should optimize the network such that only the sounding regions have a high response in S_i. Since the ground-truth sounding map is unknown, we apply differential thresholding on S_i to predict sounding objectness map $O_i = sigmoid((S_i - \epsilon)/\tau)$ [11], where ϵ is the threshold, and τ denotes the temperature that controls the sharpness.

In an egocentric video clip, a visual scene is usually temporally dynamic. Sometimes a single audio-visual pair (I_i,s) may not be audio-visually correlated. To this end, we solve the localization task in the Multiple-Instance Learning (MIL) [47] setting to improve robustness. Concretely, we use a soft MIL pooling function to aggregate the concatenated attention maps $\boldsymbol{S} = [S_1; ...; S_T]$ by assigning different weights to S_t at different time steps:

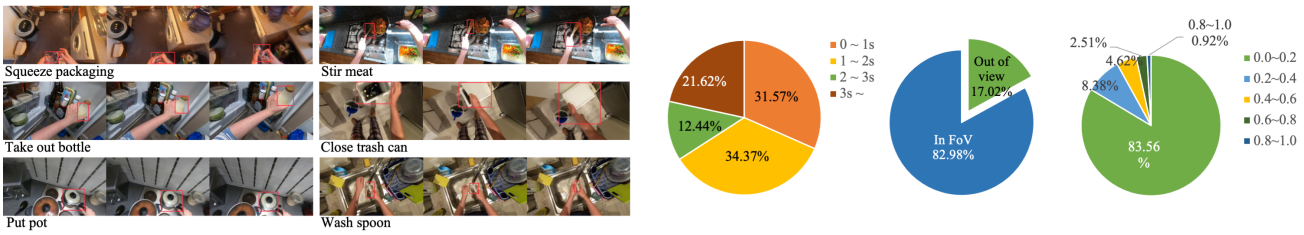
$$\overline{S} = \sum_{t=1}^{T} (W_t \cdot \boldsymbol{S})[:, :, t], \tag{9}$$

where $W_t[x, y, :] = \textsc{Softmax}(\boldsymbol{S}[x, y, :])$, x and y are the indices on spatial dimensions. Subsequently, an aggregated sounding objectness map $\overline{O}$ is calculated from $\overline{S}$. In this way, for each video clip V in the batch, we can define its positive and negative training signals as:

$$P = \frac{1}{|\overline{O}|}\langle \overline{O}, \overline{S} \rangle, \quad N = \frac{1}{hw}\langle \mathbf{1}, S_{neg} \rangle, \tag{10}$$

where $\langle \cdot, \cdot \rangle$ is the Frobenius inner product.

We obtain negative audio-visual attention maps S_{neg} by associating the current visual inputs I with audio from other



(a) Epic Sounding Object Dataset examples (b) Duration of videos (c) Num of out-of-view sounds (d) Sounding object area

Figure 5. **Illustration of the *Epic Sounding Object* dataset statistics**. **(a)** Example of our video frames and sounding object annotations. Class diversity (squeeze packaging, close trash can, put the pot, etc.) **(b)**: The distribution of untrimmed video duration. **(c)**: The number of videos that are annotated as containing out-of-view sounds. **(d)**: Distribution of bounding box areas in Epic Sounding Object dataset, the majority of boxes cover less than 20% of the image area, demonstrating the difficulty of this task.

video clips. $\mathbf{1}$ denotes an all ones tensor with shape $h \times w$. Therefore, the localization optimization objective is:

$$\mathcal{L}_{loc} = -\frac{1}{N}\sum_{k=1}^{N}[\log \frac{\exp(P_k)}{\exp(P_k)+\exp(N_k)}], \qquad (11)$$

where k is the video sample index in a training batch. The overall objective is $\mathcal{L} = \mathcal{L}_{loc} + \lambda \mathcal{L}_{dis}$, where we empirically set $\lambda = 5$ in our experiments.

4. The Epic Sounding Object Dataset

Existing sound source visual localization evaluation datasets, such as SoundNet-Flickr [65], VGG-Sound Source [11], only contain *third-person* recordings. To the best of our knowledge, there is no existing dataset that is suitable for evaluating our model. Thus, we introduce an Epic Sounding Object Dataset for egocentric audio-visual sounding object localization. Built upon the well-known Epic-Kitchens [13] dataset, we collect sounding object annotations on its action recognition test set.

Data preparation. We select 13k test videos from the Epic-Kitchens action recognition benchmark as our source data. Since these videos are not originally collected for audio-visual analysis, they vary in length, and not all of them contain meaningful sounds. To verify the videos for annotations, we conduct a two-step process: We first determine if a video is silent by checking its sound-level in decibels relative to full scale. Consequently, silent videos are filtered out to provide a meaningful data source. Second, we bin the videos by their duration and show the statistics in Fig. 5 (b). The majority last less than 2 seconds, and hence we choose to trim the center 1-second clip from each video.

After pre-processing, we obtain 5,089 videos in total for annotation. For each video, we uniformly select three frames and annotate sounding objects in the frames. We follow previous works [11, 76] to use bounding boxes to annotate the objects that emit sounds. To obtain proposals of potential sounding objects automatically, we follow Epic-Kitchens [13] to use a Mask R-CNN object detector [25] trained on MS-COCO [43] and a hand-objects detector [68] that is pretrained with 42K egocentric images [13, 42, 69].

Before voting		After voting		
Video	Frames	Video	Frames	Classes
5,089	15,267	3,172	9,196	30

Table 1. Statistics of the *Epic Sounding Object* dataset.

Annotation collection. Given the pre-processed data, we then annotate the sounding objects manually. Unlike third-person view videos, the object-sound associations in the egocentric domain are more complicated. There are two main challenges: (i) Numerous egocentric videos record wearer-environment interaction (*e.g.*, a human places a dish on the table). The object-sound associations could be dynamic, and sometimes it is hard to determine what objects are emitting sounds; and (ii) the objects in egocentric videos are often missing from the screen, resulting in variations in scale (see Fig. 5 (c)). We address the above issues by taking advantage of human commonsense knowledge. We ask three or more annotators from the Amazon Mechanic Turk to annotate the same video (frames). Concretely, they do this by first watching the 1-second video with three annotated frames to confirm what objects make sounds in the video. During the annotation course, they are asked to answer two questions: (1) Does the video contain out-of-view sounds? (2) Which bounding boxes correspond to the sounding objects? We collect the annotations in multiple rounds until each video has at least three or more valid annotations from the Amazon annotators. Finally, we conduct an annotation verification by voting on all videos. If at least two annotators agree on the same answer, it will be considered as a correct annotation; If not, we will simply omit the video. The annotation statistics after voting are shown in Tab. 1. We obtain 30 classes of sounds by counting the noun (object) classes. The annotations are evenly split into two sets for validation and testing. The examples and statistics in Fig. 5 illustrate the diverse and complicated nature of egocentric audio-visual scenes.

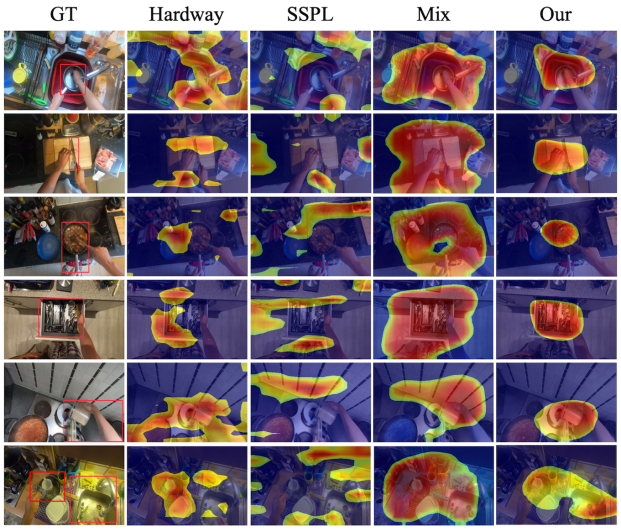


Figure 6. Qualitative comparison on *Epic Sounding Object* dataset. We show diverse sounding objects in the kitchen scenes in the first column, sounding objects are annotated in red boxes. Our method outperforms all the compared works.

5. Experiment

Datasets. In our experiments, we use two egocentric datasets. (1) Epic-Kitchens [13]: The dataset consists of 100 hours of egocentric recordings from 45 kitchen scenes. Thus, diverse kitchen-relevant events and sounds are in the dataset. We follow the same data split released in their action recognition benchmark and select 62,413 training videos. We then filter out the silent videos by measuring the Decibels relative to full scale. This results in 47,214 training videos in total. For evaluation, we use our annotated *Epic Sounding Object* dataset to report the results; (2) Ego4D [24]: Ego4D is the most recent large-scale egocentric video dataset. Besides kitchen scenes, it includes diverse daily life scenarios. Specifically, it consists of different subsets serving different benchmarks. We select the "Bristol" subset as it contains diverse scenarios (*e.g.*, entertainment, sports, commuting, and more) to test our method. We randomly sample and trim 50,000 1-second videos from this subset and use 90%/10% as the train/test split. A similar filtering strategy is applied to obtain 26,858 videos for training. We conduct experiments in Sec. 5.1.3 on this dataset to showcase the generalization ability of our method.

Evaluation metric. We follow the prior works [11, 28, 38] and adopt the pixel-level measurement for evaluating localization performance. Given the ground truth sounding object bounding boxes, we compute the Consensus Intersection over Union (CIoU) and Area Under Curve (AUC) between the predicted localization map and ground truth boxes. We report CIoU over a range of thresholds to expose the finer aspects of comparison.

Implementation details. To facilitate the training, we cut a 1-second long video around the center of each raw video. We select the middle frame from the video clip and its four neighboring frames with an interval of 2 between frames. Consequently, we get $T = 5$ frames as visual input. During training, the frames are first resized to 256×256 and then randomly cropped to 224×224. During inference, all the frames are directly resized to the desired size without cropping. For the audio stream, we extract the corresponding 1-second audio clip to create the audio-visual pairs. The audio waveform is sub-sampled at 11kHz and transformed into a spectrogram with a Hann window of size 254 and a hop length of 64. The obtained spectrogram is subsequently resampled to 128×128 to feed into the audio network. We set the number of audio and visual feature channels as 512 and choose $\epsilon = 0.5$ and $\tau = 0.03$. All models are trained with the Adam optimizer, with a learning rate of 10^{-4} on the visual encoder and temporal network, while using a learning rate of 10^{-2} for updating the audio encoder.

	CIoU			AUC
	@0.2	@0.3	@0.4	
Attention [65]	7.12	-	-	6.42
STM [38]	12.10	7.64	4.01	8.87
Hardway [11]	24.51	13.55	6.10	13.38
SSPL [71]	13.62	8.10	4.45	9.56
Mix [28]	26.01	15.25	9.90	15.39
Our	**38.71**	**19.42**	**10.51**	**18.38**

Table 2. Quantitative comparison of localization results on *Epic Sounding Object* dataset. All methods are re-trained on Epic-Kitchen. The results of metrics CIoU@{0.2, 0.3, 0.4} and AUC are reported. The top-1 results are highlighted.

5.1. Results

5.1.1 Experimental Comparison

To validate the effectiveness of our framework, we compare it with recent audio-visual localization methods: Attention [65], STM [38], Hardway [11], SSPL [71] and Mix [28]. Among all the comparative methods, STM [38] utilizes weak labels for the training, while the other methods are trained with self-supervision. We hence adjust STM [38] with our self-supervised localization loss. As all the methods are developed for third-person view videos, we retrain their methods on our training data for a fair comparison. The quantitative results are shown in Tab. 2. We can find that our method outperforms all the compared approaches by a large margin in all metrics, indicating the benefits of mitigating out-of-view sounds and explicitly modeling egomotion in learning egocentric audio-visual localization. Moreover, we provide a qualitative comparison to visually showcase our localization results. In Fig. 6, we can see that our model produces localization results that are tight around the ground truth sounding objects.

Model	TM			SL	L_{dis}	CIoU@0.2	AUC
	Avg	Max	GA				
a						27.41	15.36
b	✓					31.84	15.79
c		✓				33.29	16.10
d			✓			37.38	16.59
f			✓	✓		38.21	17.92
g			✓	✓	✓	**38.71**	**18.38**

Table 3. Ablations on GATM, SL, and audio disentanglement module. The top-1 result in each column is highlighted.

5.1.2 Ablation Study

We conduct an ablation study to illustrate how each module affects localization performance. As shown in Tab. 3, we compare our full model with different baselines — **model a**: we remove all the modules and only use the features from the visual and audio encoders to compute the localization map; **model b-d**: we insert *Average*, *Max*, and *Geometry-Aware* Temporal modeling approaches separately in the framework; in **model f**, we incorporate the soft localization (SL) into the pipeline; and in **model g**, we employ the audio disentanglement module and train the model with L_{dis}. By comparing **a** and **b-c**, we found that it's crucial to aggregate temporal context, while **d** emphasizes the importance of GATM in mitigating the egomotion in egocentric videos. **model d** vs. **f** shows that SL slightly enhances the performance since some of the unrelated visual content can be reweighted. The comparison between **f** and **g** demonstrates that by incorporating audio feature disentanglement, the localization performance can be further boosted because it can handle the out-of-view sounds in videos.

Naive baseline. To assess the difficulty of the task, we provide a center box method that predicts a gaussian heatmap around the center. This results in a naive baseline of **16.51** compared to **27.41 (model a)** from Tab. 3, showing that there are various challenging scenarios apart from the object being in the center that this naive baseline cannot capture.

5.1.3 Generalization to More Scenarios

The experiments on *Epic Sounding Obejct* dataset demonstrate the effectiveness of our method in localizing sounding objects in the egocentric videos. To further validate the generalization ability of our method, we train our audio-visual sounding object localization network on Ego4D [24] and qualitatively showcase the localization results in Fig. 7. The examples are all selected from the Ego4D test set. We can see that our model can learn audio-visual associations and localize the sounding objects in diverse scenes.

6. Discussions and Conclusions

In this work, we tackle a fundamental task: egocentric audio-visual localization to promote the field of study in egocentric audio-visual video understanding. The uniqueness of egocentric videos, such as egomotions and out-of-view sounds pose significant challenges to learning fine-grained audio-visual associations. To address these problems, we propose a new framework with a cascaded feature enhancement module to disentangle visually indicated audio representations and a geometry-aware temporal modeling module to mitigate egomotion. Extensive experiments on our annotated *Epic Sounding Object* dataset underpin the findings that explicitly mitigating out-of-view sounds and egomotion can boost localization performance and learn the better audio-visual association for egocentric videos.

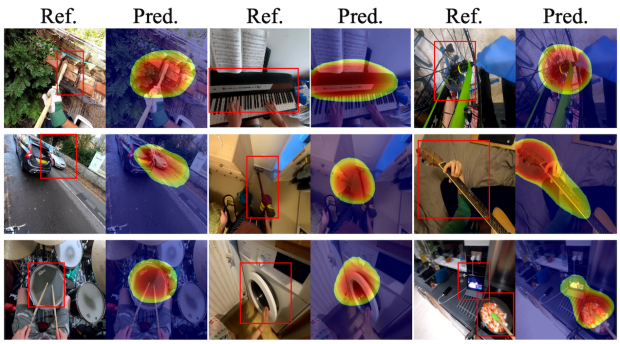


Figure 7. Localization results on diverse scenarios in Ego4D [24]. Ref.: Sounding objects; Pred.: predicted localization results.

Limitations. The proposed geometry-aware temporal modeling approach requires geometric transformation computation. For certain visual scenes with severe illumination changes or drastic motions, the homography estimation may fail. Then, our GATM will degrade to a vanilla temporal modeling approach. To mitigate the issue, we can consider designing a more robust geometric estimation approach.

Potential Applications. Our work offers potential for several applications: (a) Audio-visual episodic memory. As egocentric video records what and where of an individual's daily life experience, it would be interesting to build an intelligent AR assistant to localize the object ("*where did I use it?*") by processing an audio query, *e.g.*, an audio clip of "vacuum cleaner"; (b) Audio-visual object state recognition. In egocentric research, it is important to know the state of objects that human is interacting with, while the human-object interaction often makes a sound. Therefore, localizing objects by sounds provides a new angle in recognizing an object state; (c) Audio-visual future anticipation: following the audio-visual object state recognition task, it's natural to predict the trajectory of a sounding object by analyzing the most recent audio-visual clips.

Acknowledgements: This work was supported by Meta Research. The article solely reflects the opinions and conclusions of its authors but not the funding agents.

References

[1] Yazan Abu Farha, Alexander Richard, and Juergen Gall. When will you do what?-anticipating temporal occurrences of activities. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5343–5352, 2018. 2

[2] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. Self-supervised learning of audio-visual objects from video. In *European Conference on Computer Vision*, pages 208–224. Springer, 2020. 2, 5

[3] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 609–617, 2017. 2

[4] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *Proceedings of the European conference on computer vision (ECCV)*, pages 435–451, 2018. 2, 5

[5] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. *Advances in neural information processing systems*, 29, 2016. 2

[6] David A Bulkin and Jennifer M Groh. Seeing sounds: visual and auditory interactions in the brain. *Current opinion in neurobiology*, 16(4):415–419, 2006. 2

[7] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015. 2

[8] Minjie Cai, Kris M Kitani, and Yoichi Sato. Understanding hand-object manipulation with grasp types and object attributes. In *Robotics: Science and Systems*, volume 3. Ann Arbor, Michigan;, 2016. 2

[9] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 2

[10] Alejandro Cartas, Jordi Luque, Petia Radeva, Carlos Segura, and Mariella Dimiccoli. How much does audio matter to recognize egocentric object interactions? *arXiv preprint arXiv:1906.00634*, 2019. 2

[11] 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, 5, 6, 7

[12] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 720–736, 2018. 1, 2

[13] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision. *arXiv preprint arXiv:2006.13256*, 2020. 1, 2, 6, 7

[14] Dima Damen, Teesid Leelasawassuk, and Walterio Mayol-Cuevas. You-do, i-learn: Egocentric unsupervised discovery of objects and their modes of interaction towards video-based guidance. *Computer Vision and Image Understanding*, 149:98–112, 2016. 2

[15] Jacob Donley, Vladimir Tourbabin, Jung-Suk Lee, Mark Broyles, Hao Jiang, Jie Shen, Maja Pantic, Vamsi Krishna Ithapu, and Ravish Mehra. Easycom: An augmented reality dataset to support algorithms for easy communication in noisy environments. *arXiv preprint arXiv:2107.04174*, 2021. 1

[16] 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. *arXiv preprint arXiv:1804.03619*, 2018. 2

[17] Alireza Fathi, Xiaofeng Ren, and James M Rehg. Learning to recognize objects in egocentric activities. In *CVPR 2011*, pages 3281–3288. IEEE, 2011. 1

[18] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 5

[19] Antonino Furnari and Giovanni Maria Farinella. Rolling-unrolling lstms for action anticipation from first-person video. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):4021–4036, 2020. 2

[20] Chuang Gan, Deng Huang, Hang Zhao, Joshua B Tenenbaum, and Antonio Torralba. Music gesture for visual sound separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10478–10487, 2020. 2

[21] 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. 2

[22] Ruohan Gao and Kristen Grauman. 2.5 d visual sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 324–333, 2019. 2

[23] Ruohan Gao and Kristen Grauman. Co-separating sounds of visual objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3879–3888, 2019. 2, 4

[24] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. *arXiv preprint arXiv:2110.07058*, 2021. 2, 7, 8

[25] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 6

[26] 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

[27] 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, 2020. 2

[28] Xixi Hu, Ziyang Chen, and Andrew Owens. Mix and localize: Localizing sound sources in mixtures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10483–10492, 2022. 2, 7

[29] Robert A Jacobs and Chenliang Xu. Can multisensory training aid visual learning? a computational investigation. *Journal of vision*, 19(11):1–1, 2019. 2

[30] Hao Jiang and Kristen Grauman. Seeing invisible poses: Estimating 3d body pose from egocentric video. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3501–3509. IEEE, 2017. 2

[31] J Adam Jones, J Edward Swan, Gurjot Singh, Eric Kolstad, and Stephen R Ellis. The effects of virtual reality, augmented reality, and motion parallax on egocentric depth perception. In *Proceedings of the 5th symposium on Applied perception in graphics and visualization*, pages 9–14, 2008. 1

[32] Kazuhiko Kawamura, A Bugra Koku, D Mitchell Wilkes, Richard Alan Peters, and Ali Sekmen. Toward egocentric navigation. *International Journal of Robotics and Automation*, 17(4):135–145, 2002. 1

[33] 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. 1, 2

[34] Daekyum Kim, Brian Byunghyun Kang, Kyu Bum Kim, Hyungmin Choi, Jeesoo Ha, Kyu-Jin Cho, and Sungho Jo. Eyes are faster than hands: A soft wearable robot learns user intention from the egocentric view. *Science Robotics*, 4(26):eaav2949, 2019. 1

[35] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. *Advances in Neural Information Processing Systems*, 31, 2018. 2

[36] Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman. Discovering important people and objects for egocentric video summarization. In *2012 IEEE conference on computer vision and pattern recognition*, pages 1346–1353. IEEE, 2012. 1, 2

[37] Cheng Li and Kris M Kitani. Pixel-level hand detection in ego-centric videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3570–3577, 2013. 1

[38] Sizhe Li, Yapeng Tian, and Chenliang Xu. Space-time memory network for sounding object localization in videos. *arXiv preprint arXiv:2111.05526*, 2021. 2, 7

[39] Yin Li, Alireza Fathi, and James M Rehg. Learning to predict gaze in egocentric video. In *Proceedings of the IEEE international conference on computer vision*, pages 3216–3223, 2013. 1

[40] Yin Li, Miao Liu, and James M Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. In *Proceedings of the European conference on computer vision (ECCV)*, pages 619–635, 2018. 1

[41] Yanghao Li, Tushar Nagarajan, Bo Xiong, and Kristen Grauman. Ego-exo: Transferring visual representations from third-person to first-person videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6943–6953, 2021. 1, 2

[42] Yin Li, Zhefan Ye, and James M Rehg. Delving into egocentric actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 287–295, 2015. 6

[43] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 6

[44] 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

[45] Miao Liu, Siyu Tang, Yin Li, and James M Rehg. Forecasting human-object interaction: joint prediction of motor attention and actions in first person video. In *European Conference on Computer Vision*, pages 704–721. Springer, 2020. 2

[46] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. 5

[47] Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. *Advances in neural information processing systems*, 10, 1997. 5

[48] Roberto Martin-Martin, Mihir Patel, Hamid Rezatofighi, Abhijeet Shenoi, JunYoung Gwak, Eric Frankel, Amir Sadeghian, and Silvio Savarese. Jrdb: A dataset and benchmark of egocentric robot visual perception of humans in built environments. *IEEE transactions on pattern analysis and machine intelligence*, 2021. 1

[49] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2630–2640, 2019. 2

[50] Himangi Mittal, Pedro Morgado, Unnat Jain, and Abhinav Gupta. Learning state-aware visual representations from audible interactions. In *Proceedings of the European conference on computer vision (ECCV)*, 2022. 2

[51] Shentong Mo and Pedro Morgado. A closer look at weakly-supervised audio-visual source localization. *arXiv preprint arXiv:2209.09634*, 2022. 2

[52] Shentong Mo and Pedro Morgado. Localizing visual sounds the easy way. *arXiv preprint arXiv:2203.09324*, 2022. 2

[53] Francesca Morganti, Stefano Stefanini, and Giuseppe Riva. From allo-to egocentric spatial ability in early alzheimer's disease: a study with virtual reality spatial tasks. *Cognitive neuroscience*, 4(3-4):171–180, 2013. 1

[54] Jonathan Munro and Dima Damen. Multi-modal domain adaptation for fine-grained action recognition. In *Proceed-*

ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 122–132, 2020. 1

[55] Tushar Nagarajan, Christoph Feichtenhofer, and Kristen Grauman. Grounded human-object interaction hotspots from video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8688–8697, 2019. 2

[56] Evonne Ng, Donglai Xiang, Hanbyul Joo, and Kristen Grauman. You2me: Inferring body pose in egocentric video via first and second person interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9890–9900, 2020. 2

[57] Curtis Northcutt, Shengxin Zha, Steven Lovegrove, and Richard Newcombe. Egocom: A multi-person multi-modal egocentric communications dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 1

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

[59] Andrew Owens, Jiajun Wu, Josh H McDermott, William T Freeman, and Antonio Torralba. Ambient sound provides supervision for visual learning. In *European conference on computer vision*, pages 801–816. Springer, 2016. 2

[60] Hyun Soo Park, Jyh-Jing Hwang, Yedong Niu, and Jianbo Shi. Egocentric future localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4697–4705, 2016. 1

[61] Ivan Poupyrev, Tadao Ichikawa, Suzanne Weghorst, and Mark Billinghurst. Egocentric object manipulation in virtual environments: empirical evaluation of interaction techniques. In *Computer graphics forum*, volume 17, pages 41–52. Wiley Online Library, 1998. 1

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

[63] Andrew Rouditchenko, Hang Zhao, Chuang Gan, Josh McDermott, and Antonio Torralba. Self-supervised audio-visual co-segmentation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2357–2361. IEEE, 2019. 2

[64] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 5

[65] 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*, pages 4358–4366, 2018. 2, 5, 6, 7

[66] Silvia Serino, Francesca Morganti, Fabio Di Stefano, and Giuseppe Riva. Detecting early egocentric and allocentric impairments deficits in alzheimer's disease: An experimental study with virtual reality. *Frontiers in aging neuroscience*, 7:88, 2015. 1

[67] Ladan Shams and Aaron R Seitz. Benefits of multisensory learning. *Trends in cognitive sciences*, 12(11):411–417, 2008. 2

[68] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9869–9878, 2020. 2, 6

[69] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Charades-ego: A large-scale dataset of paired third and first person videos. *arXiv preprint arXiv:1804.09626*, 2018. 1, 2, 6

[70] Krishna Kumar Singh, Kayvon Fatahalian, and Alexei A Efros. Krishnacam: Using a longitudinal, single-person, egocentric dataset for scene understanding tasks. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9. IEEE, 2016. 2

[71] 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, 7

[72] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 2

[73] Charles Spence and Sarah Squire. Multisensory integration: maintaining the perception of synchrony. *Current Biology*, 13(13):R519–R521, 2003. 2

[74] Yu-Chuan Su and Kristen Grauman. Detecting engagement in egocentric video. In *European Conference on Computer Vision*, pages 454–471. Springer, 2016. 2

[75] J Edward Swan, Adam Jones, Eric Kolstad, Mark A Livingston, and Harvey S Smallman. Egocentric depth judgments in optical, see-through augmented reality. *IEEE transactions on visualization and computer graphics*, 13(3):429–442, 2007. 1

[76] 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*, pages 2745–2754, 2021. 2, 6

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

[78] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 247–263, 2018. 2

[79] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in the wild. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition workshops*, 2019. 2

[80] 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. 5

[81] Sudheendra Vijayanarasimhan, Susanna Ricco, Cordelia Schmid, Rahul Sukthankar, and Katerina Fragkiadaki. Sfm-

net: Learning of structure and motion from video. *arXiv preprint arXiv:1704.07804*, 2017. 5

[82] Xiaohan Wang, Linchao Zhu, Heng Wang, and Yi Yang. Interactive prototype learning for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8168–8177, 2021. 1

[83] Yu Wu and Yi Yang. Exploring heterogeneous clues for weakly-supervised audio-visual video parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1326–1335, 2021. 2

[84] Yu Wu, Linchao Zhu, Yan Yan, and Yi Yang. Dual attention matching for audio-visual event localization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6292–6300, 2019. 2

[85] Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer. Audiovisual slowfast networks for video recognition. *arXiv preprint arXiv:2001.08740*, 2020. 2

[86] Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. Dilated residual networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 472–480, 2017. 3

[87] 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. 2, 4

[88] Hang Zhou, Xudong Xu, Dahua Lin, Xiaogang Wang, and Ziwei Liu. Sep-stereo: Visually guided stereophonic audio generation by associating source separation. In *European Conference on Computer Vision*, pages 52–69. Springer, 2020. 2

[89] Yipin Zhou and Tamara L Berg. Temporal perception and prediction in ego-centric video. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4498–4506, 2015. 2